Vocal Tract Length Perception and the Evolution of Language

by

William Tecumseh Sherman Fitch III

B.A. Biology, Brown University, 1986

Thesis

Submitted in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy in the Department of
Cognitive and Linguistic Sciences at Brown University

May 1994

This dissertation by William Tecumseh Sherman Fitch III
is accepted in its present form by the
Department of Cognitive and Linguistic Sciences
as satisfying the dissertation requirement for the degree of
Doctor of Philosophy

Date..... .....
Philip Lieberman

Recommended to the Graduate Council

Date..... .....
Peter D. Eimas

Date..... .....
Marc D. Hauser

Approved by the Graduate Council

Date..... .....

Abstract

The length of the vocal tract is correlated with body size and determines the overall dispersion of formant frequencies in speech. In this thesis I explore the interconnections between vocal tract length, formant dispersion and perceived body size. I used computer-synthesized vowel sounds to show that human subjects use vocal tract length (along with other cues) to gauge the relative body size of a speaker. Vocal tract length assessment may play an important role in "vocal tract normalization", which is crucial for speech perception and language acquisition: listeners must adjust for size differences between speakers' vocal tracts to accurately perceive speech.

A second set of experiments used synthesized vowels to show that the lengthening of the vocal tract accompanying the production of [u] ("boot") vs. [i] ("beet") is perceived by human listeners as an increase in apparent size. This may explain the long-noted phenomenon of "phonetic symbolism" for size in many languages: diminutives and words for small things contain [i], while words for "large" contain [o] or [u], far more often than predicted by chance. A control experiment showed that the phenomenon is not due to other acoustic factors or to conventional linguistic associations.

Non-human primates may also use formant dispersion as a cue to body size. If animals use vocal tract length to gauge the size of a vocalizer, we might expect them to lengthen the vocal tract in situations where it would be advantageous to seem larger (e.g., during aggressive interactions). Video analysis of captive saki monkeys (*Pithecia pithecia*) showed that they used lip protrusion, which lengthens the vocal tract, during behaviors accompanied by fur erection, which increases their visually-apparent size. These data, along with much of the bioacoustics literature, suggest that non-humans also use vocal tract length as a cue to body size, and may explain the origin of the primate "fear grimace" and human smile. The framework developed in this thesis provides a unifying theoretical structure within which to empirically investigate important relationships between animal communication systems and human language.

Acknowledgments

I offer my profound thanks to:

The members, past and present, of the Department of Cognitive and Linguistic Sciences at Brown University, with special thanks to Paul Allopenna, Amit Almor, Jean Andruski, Katherine Demuth, Steve Finney, Nelson Francis, Polly Jacobson, Mark Johnson, Harriet Magen, Jim Morgan, Thannassis Protopapas, David Sheinberg and Jennifer Utman, conversations with whom helped shaped the ideas that went into this thesis.

Phil Lieberman for being a thoughtful and enthusiastic advisor, Caroline Wiltshire for her careful and insightful editorial comments, Marc Hauser for helping to sharpen my evolutionary arguments, and Peter Eimas for being helpful and open-minded, despite our differences of opinions, and for excellent advice during Phil's travels.

Hannelore Zoller and Natasha Ignatieff Fitch for their crucial role in inspiring my original interest in language.

Liza Bakewell, Judith Myers Fitch, Gray Fitch Scariot and Jana Signe for illuminating conversations and welcome respites from academia.

Jonathan Fritz for dragging me to the zoo to watch sakis the first time, and for his incisive criticisms throughout the project.

My cousin, William Wiggins Fitch, for his insights into matters poetic and metaphoric.

Eric Nicolas for his constant friendship, unswerving faith and never-failing critical eye. A friend indeed.

Cynthia Romano and Alison Bowden for being there, in more ways than I can count and with more heart than anyone could ever expect.

And most of all to my father, William Tecumseh Sherman Fitch, for his constant encouragement and support of my doing "what's right for me."

This thesis is dedicated to the memory of my sister, Hilary Durbin Fitch.

The research reported here was supported by fellowships from Brown University, the National Science Foundation, and the National Institutes of Health, and by a grant from Apple Computer, Inc.

Table of Contents

List of Tables

List of Illustrations

Introduction

> ...As for me, I am proud of my close kinship with other animals. I take a jealous pride in my Simian ancestry. I like to think that I was once a magnificent hairy fellow living in the trees, and that my frame has come down through geological time via sea-jelly and worms and Amphioxus, Fish, Dinosaurs and Apes. Who would exchange these for the pallid couple in the Garden of Eden?
>
> – W. N. P. Barbellion

As a biologist I have always found it strange that many people seek to find firm and decisive boundaries between humans and animals. Although humans are clearly different from other animals, and these differences are fascinating, the similarities seem equally interesting to me. Furthermore, all species possess idiosyncrasies and specialized adaptations, quite wonderful in their variety, which make them unique from others. But a biology which focused on differences alone would be strange indeed, given the overwhelming similarities between extant species, from the biochemical up to the behavioral level. Clearly, if we focus on minute enough, or coarse enough, details, we could enumerate the differences or the similarities between any two species for decades (indeed, the same is true of a comparison between two individuals of the same species). But making sense of nature requires that we take a view that enables us to learn which differences make a difference, and which similarities are deep homologies rather than superficial resemblances.

This thesis is a search for connections between animal communication systems and human language, for homologous mechanisms that are important in both communication systems. (For the purposes of this thesis, I define language broadly and simply as the vocal communication system used by humans, viewing such phenomena as written or signed language as recent derivatives with little importance in the evolution of language.) I believe that a comprehensive understanding of the similarities between human and animal communication systems is a necessary prerequisite for an empirical evaluation of the differences, and my search for these similarities is based firmly on the belief that humans, and the biological basis for language, evolved according to the principles set forth by Darwin (1859) in his theory of evolution by natural selection. One of the basic assumptions underlying this thesis is that the mechanisms responsible for human language evolved gradually from pre-existing mechanisms, which may well have been used for different purposes by our ancestors (traits that evolved to solve one purpose, and then were put to another, are said to have provided a "preadaptation" for the second purpose, with no connotation of foresight on the part of evolution). Given this possibility (indeed likelihood) of preadaptation, it is reasonable to wonder if the precursor of language was a communication system at all.

There is a good reason to expect that most of the mechanisms that underlie human language derived from earlier mechanisms concerned with communication. Vocal communication is a general characteristic of mammals, and it is clear that our pre-linguistic hominid ancestors communicated vocally. Surprisingly complex and specific features of both sound production and perception are shared between humans and other mammals, and these were probably used in many cases to accomplish the same general tasks: finding and evaluating mates, maintaining social cohesion, recognizing and intimidating intruders, warning of danger, advertising food, and locating and guiding dependent young. Virtually everything currently known about the neural, physiological and anatomical basis for the production and perception of speech indicates that it is shared between humans and animals.

Because modern human language makes use of the same basic means as other mammalian vocal communication systems (i.e., the same ears and basic auditory system, the same vocal tract and articulators, and the same basic neuroanatomy), an evolving language had to coexist (at the very least) with the pre-existing vocal communication system. Because of the importance of the behaviors listed above to survival and reproduction, a hypothetical mutant in which a new language mechanism replaced some previously existing mechanism crucial to the communication system of its species would

find itself less able to survive and reproduce, and would thus be selected against. Given the high degree of functional overlap which we can infer between the two systems, it seems likely that many mechanisms underlying language evolved by gradually modifying perceptual and productive mechanisms already available as part of a pre-existing communication system.

This is not to suggest that all aspects of human language derived from vocal (or communicative) sources. Human language appears to differ in some fundamental ways from other species' communication systems, and it seems rather likely that some of the mechanisms responsible for these differences were preadaptations from an entirely different domain. It seems plausible, for example, that some of the mechanisms underlying the abstract combinatoric power of language and the ability to make infinite use of finite means came from sources having nothing specifically to do with a pre-existing communicative system. However, this possibility should not obscure the fact that much of the neural and behavioral support for human language, and many of the uses to which language is put, are the same as those of any other primate communication system.

These considerations lead us to expect areas of evolutionary continuity between modern language and the pre-linguistic vocal communication system of our ancestors. Unfortunately, no traces remain of this ancient system, and we must infer its nature from studies of the present-day communication systems of humans and other mammals. These other species have been evolving in the meantime: they do not provide static copies of an earlier stage of our evolution. Even our closest relatives, the great apes (chimpanzees, gorillas and orangutans) show significant differences in their respective communication systems. To the skeptic, this might suggest that the study of the evolution of language is doomed to failure. Fortunately, however, evolution is slow, and tends to be conservative. Given the short amount of time since the divergence between humans and apes (about 5 million years, brief by evolutionary standards), there are almost certainly many aspects of the human communication system that have showed little appreciable subsequent change. The identification and characterization of these aspects of language is one of the primary goals of the biological study of language.

This thesis is an exploration of a phenomenon that appears to provide a rich and multi-leveled link between animal communication systems and human language: the use of formant frequencies as a cue to the body size of a vocalizer. I suggest that both humans and animals make use of the correlations between body size, vocal tract length and formant pattern, and that the interdependency of these three variables has played an important role in the evolution of mammalian vocal communication (and continues to be important in the present). Furthermore, I suggest that perceptual mechanisms originally involved in body size assessment formed the preadaptive basis for formant detection and vocal tract normalization, two phenomena at the heart of human speech perception. In addition, the same mechanisms appear to play an important role in the specification of meaning through the phenomenon of phonetic symbolism (non-arbitrary associations between sound and meaning in human language). Phonetic symbolism may play a role in language acquisition, historical linguistic change, and the expressive and poetic use of language. In general, then, I describe a phenomenon that plays a central role in both human and animal communication systems, and provide a theoretical framework for the investigation of what appears to be an area of significant continuity between animal communication and human language.

Chapter 1: Vocal tract length, formant frequencies and body size perception

In this chapter I discuss the acoustic basis of vocal communication in vertebrates. Because virtually all of what is known about the acoustics of vocalization derives from studies of human speech, I focus on the results of these studies. However, some aspects of human speech production are anomalous (e.g., the human vocal tract is differently shaped from that of any other mammal, Lieberman 1984), and the source/filter theory of human speech acoustics is not necessarily applicable to all non-human vocalizations. Thus, I will concentrate on areas where extensions to the source-filter theory may be necessary to understand the acoustic basis of non-human vocalization or human vocalizations other than speech (such as singing).

The major focus of this discussion is the acoustic aspects of vocalizations which provide a cue to body size. I suggest that the assessment of body size via acoustic cues has played an important selective role in the evolution of vocal communication since our earliest vertebrate ancestors began croaking 300 million years ago. Although this thesis focuses mainly on one particular cue to body size (vocal tract length and its acoustic correlate, formant dispersion), other acoustic variables such as fundamental frequency and call duration or loudness have also probably played an important role. In the following discussion I also call attention to these other possible acoustic cues to body size.

The structure of this chapter is as follows: first I briefly describe the acoustics of human speech production as modeled by the source/filter theory of speech production. Then I outline some of the difficulties with this theory, highlighting those shortcomings which are particularly relevant for the acoustic perception of body size. Based on this discussion, I specify five different possible acoustic cues to body size. Finally, I report the results of an experiment which demonstrates that human listeners make use of at least two of these cues when asked to judge the body size of computer-generated "speakers".

The Acoustics and Physiology of Vocalization

In outline, the call production system of all mammals consists of a larynx that converts a steady stream of air from the lungs into a series of puffs of air, a signal which is known as the glottal source. The glottal signal then travels through the supralaryngeal vocal tract, the length and shape of which determine a set of resonant frequencies. These resonant frequencies (called formant frequencies in human speech) then modify the glottal signal, enhancing some frequency components and weakening or eliminating others.

The source/filter theory of speech was first advanced in basic form by Müller (1848) and developed in its modern mathematical form by Chiba and Kajiyama (1939) and in Fant's classic "Acoustic Theory of Speech Production" (1960). The source/filter theory is based on the assumption that the glottal source and the supraglottal filter can be treated as functionally independent. Indeed, the term "formant" has come to connote a resonance peak which filters a source sound without having any important effect on the behavior of the source itself. This basic assumption has stood the test of time relatively well, and it seems clear some thirty years later that mathematical models based on this assumption provide a rather good approximation of normal adult male speech. However, it has become apparent that the assumption of independent source and filter is not as well justified for the voices of women and children, and it seems unlikely that it holds for all non-human primates. Fant and his colleagues have played an important role in pointing out the instances in which the assumption of independence is unjustified (e.g., Fant 1982, Fant and Ananthapadmanabha 1982, Rothenberg 1985).

The source/filter theory has provided the basis for an abundant proliferation of extremely powerful algorithms for speech processing over the last three decades, including techniques such as linear predictive coding (Markel and Gray 1976), and virtually all of the mathematical formalisms and signal processing techniques currently in use in the analysis and synthesis of speech share its assumptions. Fortunately, mathematical formalisms also exist to deal with cases of non-independent

source and filter (called variously "interactive" systems, "feedback" systems, or "coupled" systems), although these have received much less experimental and theoretical attention than the source/filter theory. These have been developed to explain the acoustic behavior of wind instruments (e.g., trumpets, clarinets, oboes and the like), whose operation requires a strong interdependence of source and filter, such that the properties of the filter act as a major determinant of the behavior of the source. It is thus profitable to construct a continuum of "interactivity", where systems with non-interactive source and filter like the adult male voice occupy the acoustically uncoupled end, and the strongly coupled wind instruments occupy the other. Where the sound-production systems of most vertebrates lie on this continuum remains an important empirical question.

Thus, a general theory of acoustic production in animals must leave the question of the degree of coupling between source and filter as a free parameter. Because the experimental portions of this thesis deal mostly with the analysis and synthesis of adult human male speech, I present an overview of the source filter theory. However, because the theoretical questions I wish to address concern more general issues in the evolution of vocal communication systems, my treatment is somewhat unconventional, emphasizing the common ground between coupled and uncoupled systems. There is, in any case, no need for a restatement of the source/filter theory in its ordinary form; the interested reader can consult Fant (1960) for a dense but mathematically complete development in terms of analog electronic theory, Markel and Gray (1976) for a clear mathematical treatment in terms of digital electronic theory, or Lieberman and Blumstein (1988) for an informal and intuitive description.

The lungs provide the airstream which powers mammalian calls; the larynx modulates this airstream, turning the energy into sound. This modulated airstream constitutes the "source" of phonated and aspirated speech, as well as most mammalian vocalizations. In fricatives and some other speech sounds, and in various hisses or clicks made by animals, the source is located elsewhere in the vocal tract. In either case, however, this source is then modified by the acoustic properties of the supralaryngeal vocal tract before radiating out into the environment.

Lungs

The primary function of the lungs is to exchange gases between the organism and the environment (mainly to supply the tissues with oxygen and rid them of carbon dioxide), and they are specialized to accomplish this task by having a huge surface area which is richly supplied with blood. However, the lungs also function as a reservoir of air for phonation. Although there are other ways to generate an airstream which can be used in sound production (e.g., forcing air out of or into the oral cavity via tongue movements, or pulling the larynx downward like a piston), none of these maneuvers can generate sounds of significant loudness or duration. Thus, the lungs provide the crucial initial element for most vertebrate vocalizations: a powerful and dependable source of moving air.

The lungs are elastic bodies; like a balloon, when fully inflated they try to deflate. We take advantage of this fact when we breathe: we contract the muscles of inspiration (diaphragm and intercostals) to draw in air during lung inflation, but we can simply relax the muscles and let the natural elasticity of the lungs accomplish exhalation. Certain diseases, notably emphysema, cause a decrease in lung elasticity such that exhalation must be accomplished via muscular contraction (of the abdominal muscles and the internal intercostals); patients with emphysema find the simple act of breathing to be exhausting. Because exhalation makes use of the elasticity of the lungs, it is easier to produce a powerful airstream on exhalation than on inhalation. Given this fact, it is not surprising that most vocalization occurs during the exhalation phase of breathing (termed egressive vocalization). In particular, virtually all human speech and singing occurs during exhalation. However, certain human vocalizations, such as intense crying or the laughter of some individuals, make use of ingressive vocalization, and ingressive phonation is occasionally found in speech (e.g., the word "oui" spoken in some dialects of French). Ingressive vocalizations are also found in the vocalizations of at least some other mammalian species (e.g., the chimpanzee "pant hoot" is produced with both inspiratory and expiratory phonation, Goodall 1986, p. 134). Because lung volume is

probably related to total body size, and the duration of either an ingressive or egressive vocalization will be limited by lung volume, call duration provides a potential cue to body size (see "Possible Acoustic Cues to Body Size", below).

Larynx

The larynx can be thought of as a valve atop the lungs which can prevent air flow out of the lungs (as when holding your breath) or the inward flow of foreign matter like food or water into the lungs. This latter function was probably the original purpose of the larynx (Negus 1949), and it remains a vital one today. When the larynx "valve" is fully open (as during quiet breathing) air can move freely in and out of the lungs; when it is tightly closed (as when we prepare to cough or lift a heavy object), a substantial back pressure can build up within the lungs because air flow is prevented.

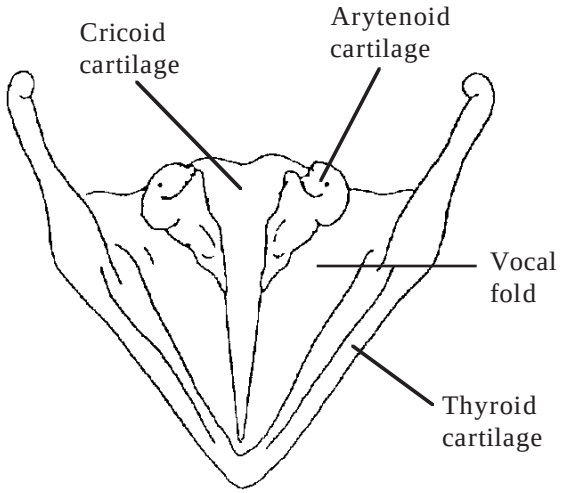


Fig. 1- View of the human larynx from above

Phonation results from a laryngeal state between these two extremes. Fig. 1 shows a view of the larynx from above: the vocal folds (i.e., "vocal cords") are stretched from front to back. The front of the folds are fixed to the thyroid cartilage, but the rear portions are attached to the mobile arytenoid cartilages, which can pivot and thus bring the folds together (adduction) or move them apart (abduction), thus changing the size of the glottis (the opening between the folds, through which air flows). If the vocal folds are loosely approximated, an increase in lung pressure will blow them apart. Then, a combination of the elasticity of the folds themselves and a Bernoulli force developed between them due to the air flow in the glottis will bring the folds back together to repeat the cycle again. This explanation of the operation of the vocal folds, first spelled out by van den Berg (1958), is known as the "myoelastic-aerodynamic theory" of phonation, and contrasts with the older ideas that phonation resulted solely from elastic forces in the vocal folds, or from a periodic tensing of the laryngeal musculature. The combination of elastic and aerodynamic forces sets up an oscillatory opening and closing of the glottis, causing what would otherwise be a steady air stream leaving the lungs to be broken into a series of puffs of air many times per second. The rate at which these puffs occur, measured in Hz (cycles per second), is known as the fundamental frequency of phonation, or F_0. This is the primary determinant of the pitch of a speech sound.

The sound which results from these puffs of air is called phonation. It is harmonic, meaning that its spectrum contains peaks of energy not only at the fundamental frequency, but also at integer multiples of that frequency. The relative amplitudes of these additional peaks (called "harmonics")

play a critical role in determining the tone quality or “timbre” of the sound. Sounds with relatively weak harmonics sound tonal and “pure” (similar to a flute) while sounds with many strong harmonics sound “buzzy” (like a harmonica). The spectrum of the human glottal source is at the "buzzy" end of this continuum, with appreciable energy in the higher harmonics.

The vocal folds are complex structures which contain intrinsic muscle fibers (the thyroarytenoid or “vocalis” muscles) and are also influenced by a number of extrinsic muscles (primarily the cricothyroids, cricoarytenoids and interarytenoids). Various combinations of tension in these muscles cause the vocal folds to behave in slightly different ways, making possible different “modes” or “registers” of phonation (van den Berg 1958, 1968). In humans, it is generally agreed that there are at least three such modes, which allow successively higher fundamental frequencies: “fry” register produces the lowest F_0s and sounds like growling or groaning, normal register is used most of the time in normal singing or speaking, and “falsetto” register sounds “thinner” and more pure and can produce the highest notes. There is considerable disagreement among speech scientists about the taxonomy and terminology of phonation registers. All of these phonation modes (described in more detail in Lieberman and Blumstein 1988), and perhaps more, are probably available to non-human mammals.

The larynx can transform the air flow from the lungs into sound not only by the mechanism of phonation described above but also via a different acoustic mechanism: turbulence. Laminar (non-turbulent) flow is exemplified by the slow movement of water in a broad, quiet river, while the irregular flow of water in a white-water stream is turbulent. Turbulence is caused when the Reynolds number of the system of air flow between the approximated vocal folds exceeds a critical value (between 2000 and 3000). Turbulence has a long-term spectrum closely approximating that of white noise (equal energy at all frequencies) and sounds similar to the static on an untuned radio. Noise produced at the larynx is called “aspiration” noise; a good example is the /h/ sound at the beginning of “Henry” (in American English). Noise can also be produced above the larynx due to a constriction elsewhere in the vocal tract, in which case it is called “frication” noise. In either case, however, turbulence is much less efficient at turning the pulmonary airstream into sound, and as a result sounds produced by aspiration or frication are typically quieter than those produced by phonation.

In birds, the sound producing source is not the larynx (which still exists, and plays its more ancient role as a simple valve into the respiratory tract), but a different vocal organ called the syrinx with no analog in mammals. The syrinx is a paired organ lying at the confluence of the two bronchial trees at the base of the trachea (thus deep in the bird's chest). The structure of the syrinx varies significantly among different bird taxa, so no detailed description of its anatomy will be attempted here. Although the details of its acoustic functioning remain unclear, the basic mechanism of its operation appears to be similar to that of the larynx: a combination of the elasticity of the syringeal membranes with Bernoulli forces caused by air rushing past them sets the membranes into oscillations. In some birds, the two syringeal membranes can vibrate independently, allowing the production of two different fundamental frequencies. The primary difference between bird and mammal vocal anatomy, for the purposes of this thesis, is that the tube filtering this syringeal source is not the supra<u>laryngeal</u> vocal tract (i.e., the pharynx and mouth) but the supra<u>syringeal</u> tract (the mouth and pharynx plus the entire trachea).

The anatomy of the larynx does not differ greatly between humans and most non-human primates, although its size does. The acoustic principles outlined above, developed from studies of humans, will probably hold reasonably well for most other primates and many mammals. The same is not necessarily true for the acoustics of the rest of the vocal tract, which we will therefore cover in more detail.

Supralaryngeal Vocal Tract

We have seen that the airstream generated by the lungs is modulated by the larynx into a series of puffs of air during phonation. This “glottal source” is further modified by the airways of the pharynx, mouth and nasal cavities, which are collectively known as the supralaryngeal vocal tract (SLVT). The SLVT contains a volume of air which has elasticity and mass, and thus can be set into vibration. An understanding of the acoustics of the vocal tract entails a detailed knowledge of these vibrations, the specification of which is the goal of mathematical models of the vocal tract. Because the mathematical basis for accurate vocal tract modeling is complex, and the equations can in practice only be solved by computer programs, I try here to develop an intuitive understanding of simpler models based on easily-understandable mathematics. The same basic principles underlie the modeling of tracts with more complex shapes.

We start with the consideration of a simple tube of uniform diameter, closed at one end (the “glottis” end) and open at the other (the “lip” end). If the diameter of the tube is small relative to the wavelength of the sounds we are interested in (which will typically be the case), propagation of energy from the glottis to the lips will occur down the length of the tube, and any spherical propagation can be ignored. The wavelength of a sound (λ) is inversely proportional to its frequency (f) according to the equation

Equation 1: $$\lambda = \frac{c}{f}$$

where c is the speed of propagation of sound in air (about 335 m/sec). Thus, the wavelength of a 100 Hz tone (a typical F_0 for a human male) is 3.35 m; because the maximum diameter of the vocal tract is on the order of 5 cm, spherical propagation can be safely ignored for this frequency. (To provide a contrast, the wavelength of an 8 kHz marmoset phee call is only about 4 cm.)

A wave of sound pressure which starts at the larynx travels down the length of the tube; when it reaches the end some of the energy will radiate out the open end, and some will be reflected back into the tube, traveling in the opposite direction. This reflected energy will travel back down the tube, ending up back at the glottis after an amount of time dependent on the length of the tube, where it will again be reflected. If another glottal pulse happened to be leaving the larynx at precisely that moment, the two pressure waves will cooperate, and the reflection will add energy to the newly-admitted pulse. If alternatively the two are out of phase, they will partially cancel. In general, then, certain frequencies will be “encouraged” by the tube, and others will be discouraged.

The frequencies which the tube "prefers" are known as its resonant frequencies, and they are determined by the length of the tube: certain wavelengths “fit” into the tube in a way consonant with its boundary conditions, while others do not. The boundary conditions are determined by the tube geometry: a closed end will prevent air movement and create a region of high pressure, while an open end will allow free movement of air and zero pressure build-up. Thus, different types of simple tubes will have different patterns of resonance.

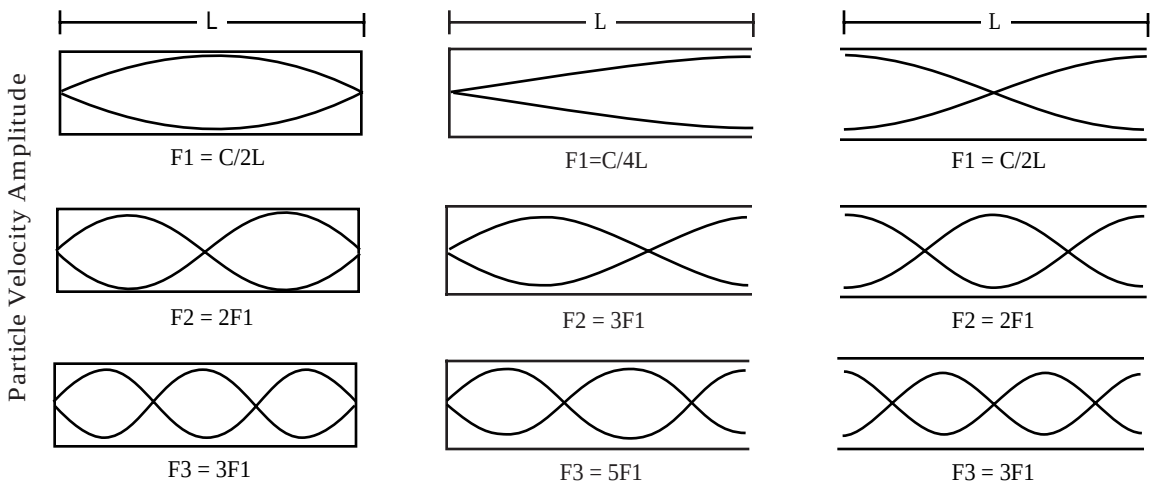


Fig. 2: First three modes of different tubes

Fig. 2 shows the first three resonances of three types of tube: one open at both ends, one open at only one end, and one closed at both ends. The tube with only one end open has a pressure maximum at one end and a zero at the other, and the lowest frequency which can meet these boundary conditions can "fit" one-fourth of its wavelength in the tube. As a result, such a tube is called a "quarter-wavelength resonator". As the diagram shows, other higher frequencies can also meet these boundary conditions, and they are related (as odd integer multiples) to the lowest resonance. The other two tubes have either highs at both ends or zeros at both ends, and thus behave as half-wave resonators: the lowest resonance has a wavelength twice the length of the tube.

To make these examples more concrete, let us calculate the first three resonances of a simplified model vocal tract 17.5 cm long (the approximate length of an adult male vocal tract). During phonation, the glottis is closed much of the time, so the quarter-wavelength resonator is the appropriate choice of tubes. The wavelength (λ) of the lowest resonance is four times the length of the tube (L = 17.5 cm), so $\lambda = 4L$. Combining this with equation 1 above we get

Equation 2: $$f = \frac{c}{\lambda} = \frac{c}{4L} = \frac{35{,}000 \text{ cm/sec}}{4 \times 17.5 \text{ cm}} = 500 \text{ Hz}$$

We can see in Fig. 2 that the next resonance is three times this frequency, i.e. $F_2 = 3c/4L = 1500$ Hz. More concisely, the entire set of resonances can be calculated from

Equation 3: $$F_{k+1} = \frac{(2k+1)c}{4L} \qquad \text{where } k = (0,1,2,3...)$$

and F_k represents the k^{th} formant. A graphical depiction of the resonances for a tube of length 17.5 cm open at one end is shown in Fig. 3; this function is known as the acoustic "transfer function" of the tube, and shows which frequencies the tube will encourage and which it will discourage. Specifically, it shows the amount of attenuation (weakening) of the input frequencies in decibels (dB), relative to the frequency which is least affected (500 Hz, in this case). Because there is a peak in the transfer function at 500 Hz, a 500 Hz sine wave input at one end of the tube will be output at the other at high amplitude. However, a 1 kHz tone falls in a valley of the transfer function, and will accordingly be heavily attenuated. It is important to realize that the transfer function is not equivalent to the output spectrum. The output spectrum represents the combination of the input signal and the transfer function,

so the spectrum of the output signal depends crucially on that of the input signal. Thus, if the input signal is a sine tone, the output spectrum will have only a single spike (revealing very little about the transfer function); if the input is white noise (with equal energy at all frequencies), the output spectrum will closely resemble the transfer function. This point is particularly important when performing acoustic analysis of high-pitched calls: the widely-spaced harmonics of the source "sample" the transfer function of the vocal tract sparsely, and thus may reveal less about it than a low-frequency or noisy source (Ryalls & Lieberman 1982).

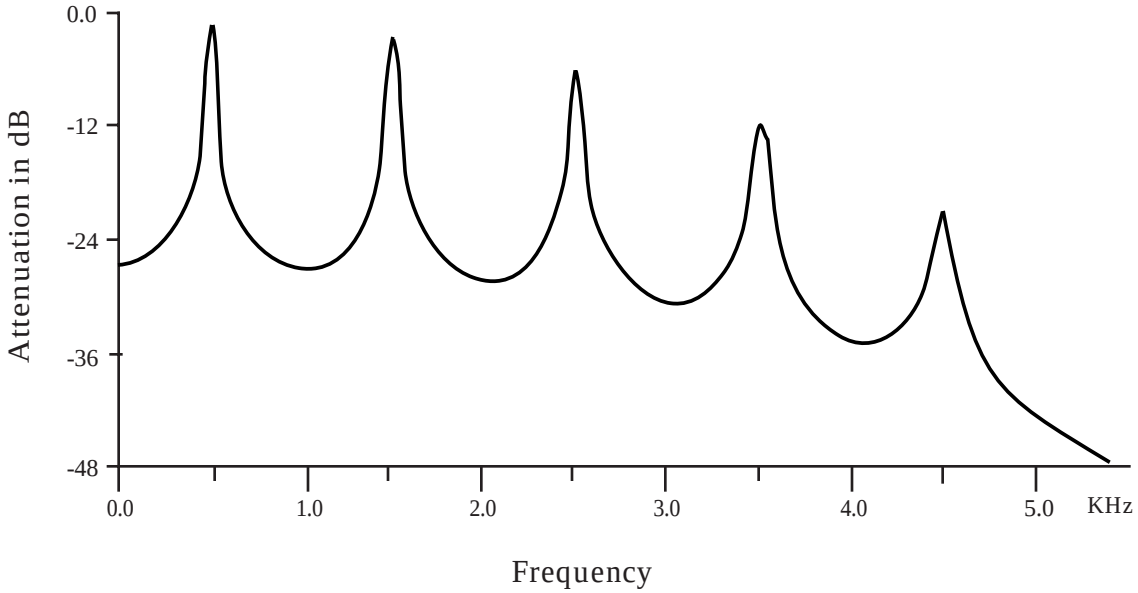


Fig. 3: Acoustic transfer function of a 17 cm tube open at one end

Understanding the acoustics of tube shapes more elaborate than those discussed above requires complex mathematics beyond the scope of this thesis (see Morse 1981, Chiba & Kajiyama 1958, Fant 1960) and calculation of the transfer function for most shapes requires computer-implemented numerical techniques. However, a few simple shapes are worth considering in passing. A conical tube, which increases diameter linearly, will possess resonances at even intervals and resembles the half-wave resonators shown in Fig. 2. This shape is seen in some wind instruments (e.g. saxophones), and can probably be approximated by mammalian vocal tracts (e.g., see Fig. 1 in Lieberman et al. 1969). Other more complex shapes which exhibit a monotonic increase in tube diameter are called "horns" and their acoustic properties depend on the exact mathematical description of their shape. "Exponential" horns are optimal for radiating energy out of the tube (such horn shapes are found in large loudspeakers), while "Bessel" horns reflect almost all of the energy back into the tube (these are found at the ends of wind instruments like trumpets). These gross differences in acoustic properties resulting from subtle shape changes should engender caution in our interpretation of primate articulatory data, since we will rarely have measurements of sufficient precision to discern such shape differences.

Although the tube models we have examined so far are obviously drastic simplifications, they illustrate a number of important points. Because we are dealing with propagation down the length of the tube, bends or curves make little acoustic difference (this is why the tubing in a French horn or tuba can be curved). Similarly, within broad limits, the diameter of the tube has little effect on its acoustical properties. The critical factor determining the transfer function of these simple tubes of constant diameter is their length. Vocal tract length is defined physically as the distance between the glottis and the lips, measured along a curve dividing the cross-sectional diameter of the vocal tract in half (see Fig. 4). Vocal tract length can be modified in two ways: it can be increased at the lip end by

tensing the orbicularis oris muscle and thus rounding and protruding the lips, or at the other end by lowering the larynx (using the laryngeal strap muscles: the sternothyroid and sternohyoid).

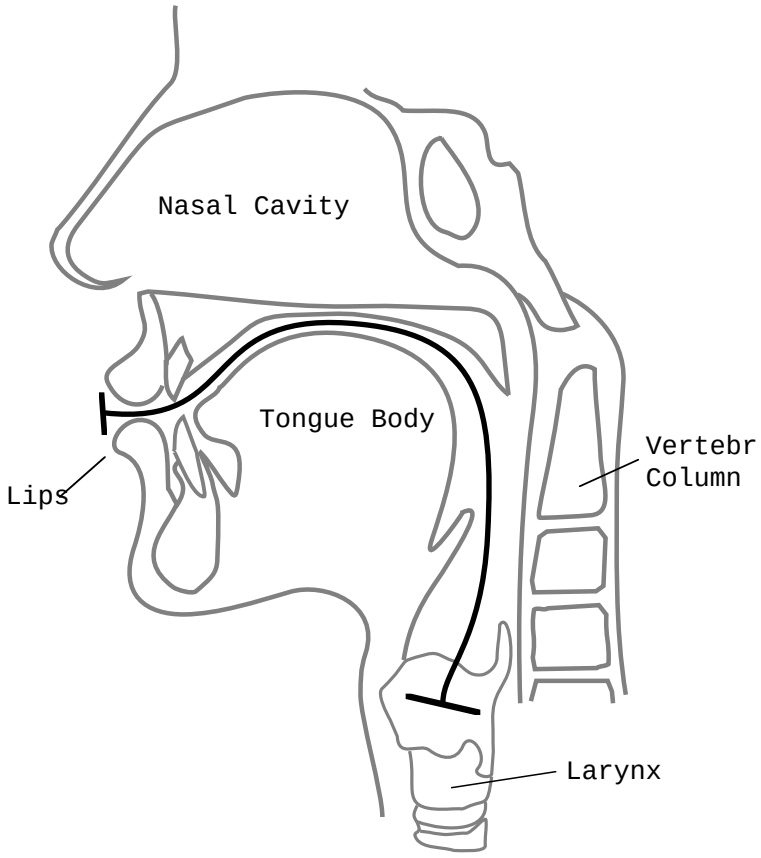


Fig. 4: Side view of the human vocal tract: the dark line illustrates the vocal tract length

The human vocal tract is not a simple tube: it has a complex shape which can be described as a series of cross-sectional area measurements, as a function of distance from the glottis. The frequencies of the formants of human speech are determined by both the length and the shape of the vocal tract. This cross-sectional area function can be modified in a variety of ways in humans. The main determinant of vocal tract shape and vowel quality is the location of the tongue body in the oral and pharyngeal cavities, which are almost at right angles to one another. A set of muscles encircling the pharynx, the pharyngeal constrictors, can shrink the pharynx, pulling the tongue backwards (as in the vowel [a] in “father”). The tongue can be pulled forward by the genioglossus muscle to make vowels like [i] (“beet”). The tongue can be pulled backwards and up in the vocal cavity to make vowels like [u] (“boot”) using the styloglossus and superior pharyngeal muscles. Opening and closing the mouth by modifying the angle of the mandible varies the size of the oral cavity accordingly.

Although the vocal tract shapes corresponding to different speech sounds are well-documented (Fant 1960, Nearey 1978, Baer et al. 1991), the muscular maneuvers necessary to achieve these shapes are poorly understood. Furthermore, all of the maneuvers described above depend on the right-angle connection between the pharyngeal and oral cavities, which is unique to humans. Thus, for non-human vertebrates, the types of muscular maneuvers available, the cross-sectional area functions possible and their acoustic consequences remain largely mysterious. Lieberman, Klatt and Wilson (1969) pioneered the study of the acoustic effects of the non-human SLVT by anatomical explorations of the range of

articulator movements in the alive (anesthetized!) macaque, and then using computer programs to model the acoustics of the observed vocal tract shapes.

In birds, as mentioned above, the pertinent filtering of the syringeal source is performed by the combination of the mouth and pharynx and the entire trachea. This, of course, makes the resonances of the bird's vocal tract considerably lower than those of a comparably-sized mammal. It also allows some interesting alterations of bird vocal tract anatomy over the course of evolution, which will be described later. The acoustics of bird vocal production have received little detailed study, and there is little agreement on even basic questions such as the role of the vocal tract in phonation. Greenewalt (1968) suggested, on the basis of the assumptions of the source / filter theory of speech perception and calculations based on the assumption of a static vocal tract, that the vocal tract plays no important role in bird vocalization. However, Nowicki (1987) showed that the vocal tract plays an active and important filtering role by studying the changes in vocalization which occurred when birds sang in a helium / oxygen atmosphere. (This change occurs because the speed of sound in helium is nearly twice that in nitrogen, resulting in an increase of formant frequencies.) Nowicki found that the modifications of vocal tract resonance frequencies caused by the helium atmosphere resulted in drastic changes in the resultant vocal output. He suggested that the nine songbird species he studied manipulate the center frequency of the vocal tract resonance so that it coincides with the fundamental or harmonics of the syringeal source. Although the question of how strongly the source and filter are coupled in bird song remains open, this result suggests that the coupling could potentially be strong.

Nasals and Nasalization

The calculation of vocal tract resonances is complicated by the potential contribution of the nasal cavities. The opening to the nasal passageways (the "velopharyngeal port") is controlled by a flap of tissue called the velum. In speech, the velum can be pulled up and back to seal off the nasal cavity, so that all air exits through the mouth. In this case only the pharyngeal and oral cavities are acoustically important. If the velum is lowered and mouth opened, air can flow through both the oral and nasal cavities: the sound is "nasalized". If the mouth is now closed, the oral cavity acts as a separate resonator attached to the nasal tube (as in [m]). Finally, the opening between the pharynx and oral cavity can be closed by bringing the tongue and velum together (as in the velar nasal [ŋ] at the end of "song"), removing the oral cavity from the acoustic circuit. Now, only the nasal tube has any acoustic effect. Because non-humans typically raise the larynx and lock it into the nasal cavity during quiet breathing, it seems likely that at least some calls are produced in the latter way, with the oral cavity playing no acoustic role.

Some evidence suggests that nasalization of vowels decreases their intelligibility (Bond 1976), leading Lieberman (1984) to suggest that an ability to produce non-nasal vowels acted as a major selective force in the evolution of the human vocal tract. However, Bond's results were obtained from English speakers, where nasalization of vowels is not distinctive. Speakers of languages where nasalization is common and linguistically-distinctive might not show such a deficit. Many languages (Portuguese is a well-known Indo-European example) use vowel nasalization as a linguistically-important distinction, and it is unlikely that use of such a feature would survive if it resulted in a large decrement in intelligibility of the language. Because 22.4 % of Maddiesson's (1984) sample of 317 languages used nasalization distinctively (more than any other secondary vowel quality, including vowel length), it seems unlikely that nasalization *per se* leads to huge deficits in intelligibility. Although as Lieberman (pers. comm.) has pointed out, many of these languages restrict nasalization to certain vowels, perhaps to limit the resulting intelligibility deficits, more data will be required before we can firmly conclude that nasalization played an important role in the evolution of the human vocal tract. In any case, however, the size of the nasal cavity influences its transfer function, and the contribution of the nasal tract to the acoustic spectrum of the output sound potentially provides another cue to body size.

Feedback and Non-Feedback systems: Source/Filter Theory of Speech

I have been deliberately vague till now regarding the precise effect of the vocal tract resonances on the glottal source. This is because there are two qualitatively different ways a system composed of larynx and vocal tract can behave. Either one (or both) may occur in the vocal repertoires of non-human species.

If the source and filter are acoustically coupled, as is expected if any of the resonant frequencies of the vocal tract are near the fundamental frequency of phonation, the two will interact: it will be much easier to maintain phonation at a resonant frequency of the vocal tract than otherwise. Such a system, where the fundamental frequency is affected or even determined by the resonant frequencies of the system, can be called a "feedback" or "coupled" system. Most wind instruments (e.g., clarinets, trumpets, oboes, tubas, etc.) work in this manner: the pitch of the instrument is controlled by the length of the resonant tube (this is particularly clear in the slide trombone). The Bessel horn found at the end of these instruments (the "bell") is optimized to reflect sound back into the tube, thus strengthening the coupling and augmenting the feedback.

In contrast, if the lowest resonances are considerably higher in frequency than F_0, as is typically the case in human speech, there will be little or no interaction between source and tract, and the F_0 of the laryngeal source will be determined solely by factors intrinsic to the larynx (e.g., the tension on, and size of, the vocal folds). This represents a "non-feedback" system, which exhibits independence between source and filter. In addition to the human voice, reed instruments with very short resonating chambers such as accordions and harmonicas fall into this category.

The source/filter theory of speech (Müller 1838, Fant 1960) takes as its starting point the independence of source and filter. For a typical male human phonating at 120 Hz, the lowest possible formant is around 250 Hz (in [u] or [i], due to the oral or pharyngeal cavities acting to a first approximation as a Helmholtz resonator), so this assumption is entirely valid. Under this assumption, the vocal tract acts simply as a filter modifying the spectrum of the glottal source, and changes in the supralaryngeal filter have no important effects on F_0 or on the glottal waveform. In this situation, the peaks in the transfer function are referred to as formants, which connotes the independence of source and filter (unlike the term resonance which suggests the possibility of their interdependence).

The source/filter theory has proven very successful as a model of the speech production process. There are some difficulties with the assumption of source/tract independence (see Klatt & Klatt 1990 for a review) but these require only minor modifications of the source/filter theory as laid out in Fant (1960), which is still widely accepted as the theory of speech production. However, given the wide range of F_0s and vocal tract lengths seen in non-human primates and other vertebrates, there is reason to suspect that the assumptions of the source/filter theory will not hold for all other animal calls. The extent to which the vocalizations of other primates (and other mammals) exhibit feedback between source and filter is a matter for empirical study; the field of bioacoustics is clearly in need of research on this topic.

Modifications of the Source/Filter Theory for Human Speech

In the classic formulation of the source/filter theory (Fant 1960), the glottal source is conceived of as depending very little on the vocal tract shape (more specifically on the vocal tract impedance). However, at least five different types of interaction between the glottal source and vocal tract impedance have been identified in human speech. The first two are specifically associated with the first formant standing wave, and the third is a result of any constriction in the vocal tract. The final two result from the fact that the terminating impedance of the glottis, which is one determinant of formant frequency, changes through the glottal cycle as the vocal cords open and close.

The glottal source can be analyzed in terms of two variables: the time-varying glottal area, and the pressure drop across the glottis. If source and filter were independent, the pressure drop would result solely from the sub-glottal pressure reserve created by the lungs. However, standing waves set up in the vocal tract (typically associated with the first formant and occurring in the pharynx) can cause an increase and decrease in the supraglottal pressure, and thus a fluctuation in the net pressure drop. Even if the mechanical behavior of the glottis is unaffected by this fluctuation, it will have an acoustic effect: the glottal waveform itself will fluctuate at the frequency of this standing wave. This nonlinear effect, described in Fant (1982) and Klatt and Klatt (1990) as "F1 ripple" in the source waveform, has the net effect of boosting the amplitude of the first formant, because energy at the frequency of F1 is already represented in the source.

A second effect associated with the first formant standing wave relates directly to the interactive source/filter theory developed by Benade (1990) for the clarinet: the F1 standing wave can be coupled to the mechanical behavior of the vocal folds such that it affects the rate or strength of vocal fold closing and opening when F1 is at or near an integer multiple of F_0 (Fant and Ananthapadmanabha 1982, Rothenberg 1985). More research is still necessary to determine if this effect plays an important role in normal speech, but there is some suggestive evidence that professional soprano singers (who frequently sing notes with fundamental frequencies at or above typical F1 frequencies) may utilize such a nonlinear source-tract interaction to aid their singing: they vary the first formant (by varying mouth opening) to coincide with F_0 (Sundberg 1975, 1987). Sundberg also found that the formant values in his singer changed abruptly when she closed the glottis completely (when instructed to hold her breath: Sundberg 1975 p. 91), suggesting coupling between supra- and sub-glottal systems. More evidence for such coupling was presented by Schutte and Miller (1986), who also noted that many experienced sopranos feel that "a well-produced tone offers an increased resistance to breath pressure", as would be expected if there were increased coupling between source and filter. Although there is thus some suggestive evidence that source/tract interactions might play an important role in the singing of high notes, neither Schutte and Miller nor Sundberg explicitly considered this possibility.

The two phenomena described above depend on the presence of a standing wave in the vocal tract, which takes some time to build up. As a result, no interaction is possible at the onset of phonation, and as the F1 standing wave develops, the interactive effect (i.e., F1 ripple in the source, amplitude and/or bandwidth of F1, and pitch shift) should "bloom". A second type of source/tract interaction would not vary in this manner: the increased impedance caused by a narrow constriction in the vocal tract (e.g., during frication). This was studied by Bickley and Stevens (1986) by applying sudden constrictions to the vocal tract during normal speech. Their results suggest that interactions occur only when the area of the constriction is quite small (comparable to that of a fricative), and that the effect is thus negligible during the production of vowels or sonorant consonants.

A fourth example of a source-tract interaction is best-viewed as an example of the source influencing the filter. The source/filter theory views the vocal tract transfer function as a static linear filter (actually, the filter changes, but at a much slower rate than the source). This view is based on the assumption that the glottal impedance (which forms one end of the vocal tract) is very high relative to the rest of the vocal tract, an assumption which is indeed true during the closed portion of the glottal waveform. However, when the glottis is open, the transfer function changes because it is based on the impedance of the entire system. The main effects of this are a significant increase in the bandwidth of F1 (Nord et al., 1986), and in the frequency of F1 (Klatt & Klatt 1990). These rapidly time-varying changes in the filter are dependent upon the frequency of phonation and the open quotient (the percentage of the glottal cycle during which the glottis is open).

Finally, during the open portion of the phonation cycle, the air column contained in the trachea and bronchi (the subglottal system) can be coupled to the SLVT, modifying the spectral effects of the supralaryngeal filter in a complex way (Fant 1972, Klatt & Klatt 1990). The acoustic properties of this subglottal system can be calculated in much the same way as outlined above for the supraglottal

system: it is a tube of air which possesses resonant frequencies dependent on its area function and total length. Because these frequencies provide a potential cue to body size, they will be discussed in more detail later (see "Possible Acoustic Cues for Body Size", below).

Summary: The Acoustics and Physiology of Vocalization

The call production system of primates and other mammals consists of a larynx which converts a steady stream of air from the lungs into a series of puffs of air, a signal which is known as the glottal source. No important differences between the larynges of humans and other primates are known, except for differences in overall size, and our current understanding of human laryngeal behavior will probably be applicable to other primates. The glottal signal then travels through the supralaryngeal vocal tract , the length and shape of which determine a set of resonant frequencies (its absolute volume and overall curvature are unimportant). The vocal tract can modify the glottal signal in one of two ways. If the resonances of the vocal tract are near the fundamental frequency of the source, F_0 will be influenced by the vocal tract resonances. This could be called a “feedback” vocal tract, and its acoustic description is similar to that of most wind instruments. Alternatively, if the resonances are considerably higher in frequency than the F_0 of the glottal source, the source and tract will be essentially independent, and the vocal tract will simply act as a filter which lets energy pass through at its resonances (called “formants” in this case). The result is that F_0 will be determined exclusively by forces intrinsic to the larynx or lungs. Such a “non-feedback” model is typified by the source/filter theory of speech, and holds relatively well for most speech as well as some wind instruments (e.g. harmonica). However, there are several difficulties with the source/filter theory, even for human vocalizations, and it is likely that some or all of these difficulties also apply for the vocalizations of other animals. Thus, in the discussion of possible acoustic cues to body size below, I will include both those predicted by the source/filter theory and those which follow from more general acoustic principles.

Possible Acoustic Cues to Body Size

There are of course two questions concerning acoustic cues to body size: does a particular cue (as measured instrumentally) actually correlate with body size (a veridical cue), and does a cue influence the size judgments made by human listeners (a perceptual, and potentially illusory, cue). From the point of view of the evolution of language, these questions are equally important, since even inaccurate perceptual mechanisms can play a powerful selective role in the evolution of a communicative system. Unfortunately, there is very little research on either question. Thus the following discussion relies heavily on acoustic theory, and I will be restricted to suggesting a number of potential cues to body size. Much more research is clearly required in this field.

Pulmonary Cues

The volume of air contained by the lungs determines the volume of air available for phonation, which in turn determines (all other things being equal) the maximum possible duration of a single vocalization. “All other things” in this case refers specifically to mode of phonation, because different modes consume stored air at different rates (van den Berg 1968). Because lung volume is closely related to thoracic volume and hence allometrically-coupled to total body size (see comments by Klatt & Klatt 1990, p. 833 and Fant et al. 1972, p. 11), call duration might serve as a cue to body size. Although I know of no data in humans or animals which bear directly upon this question, Clutton-Brock and Albon (1979) found that the duration of bouts of calling was related to fighting ability in red deer stags.

The lungs, bronchi and trachea contain air which can vibrate in a manner reminiscent of that in a Helmholtz resonator. However, due to the fact that the lungs are not hollow but spongy, there is significant internal friction resulting in extremely high damping. In any case, if the subglottal respiratory tract is acoustically coupled to the vocal tract (for example when the glottis is open), the

acoustics of this subglottal system can contribute to the overall filter function. In humans, these subglottal poles and zeroes come into play during aspiration (noise produced by a partially-open glottis, see below), and acoustic analyses have provided evidence of strong poles or subglottal formants with frequencies of around 1650 and 2350 Hz, along with a weaker spectral peak at 650 Hz for some females speakers (Klatt & Klatt 1990). Fant, Ishizaka, Lindqvist and Sundberg (1972) obtained similar values by measuring subglottal impedance in tracheotomized subjects. Because of the correlation between lung size and body size mentioned above, the frequencies of these subglottal formants provide another potential cue to body size. Such cues would be particularly evident in vocalizations with a large amount of aspiration (e.g., growling).

An interesting acoustic correlate of body size has recently been demonstrated in the bicolor damselfish *Pomacentrus partitus*. The males of this species use "chirp" sounds in behavioral interactions, and males are able to distinguish individual identity on the basis of this cue (Myrberg & Riggio 1985, Myrberg et al. 1986). It appears that the size of the fish's gas bladder determines the peak frequency of these chirps, and peak frequency is thus negatively correlated with body size in this species (Myrberg et al., 1993).

Laryngeal Cues

It seems intuitively obvious that lower pitch (which presumably means lower fundamental frequency) is a correlate of a larger body, and there is evidence from a number of species indicating that this is the case. However, the data for humans are mixed, and the only data which reliably supports this idea is that, on average, human males have lower pitched voices and are larger than females. However, no statistically significant within-gender correlation between body weight or height (or even surface area) and voice F_0 has been demonstrated to date, and numerous studies have failed to find a correlation.

Lass and Brown (1978) and Künzel (1989) both failed to find significant correlations between body weight, height or surface area and speaking fundamental frequency. In the Lass and Brown study, height and weight were measured, and body surface area estimated using a standard metric; in the Künzel study, height and weight values were those reported by the subjects themselves (no surface area data were given). In either case, there was no correlation between any of these values and average F_0.

In terms of the perception of body size from speech, the results are somewhat more promising. Lass and Davis (1976) found that 18 of 30 listeners were able to assign unseen tape-recorded speakers to weight classes accurately. Lass et al. (1978) reported that listeners were able to guess the weights of speakers to within 3.5 lbs. However, Cohen et al. (1980) showed that this claim was erroneous, being based upon averaged guesses of multiple subjects about average weights of speakers. In fact, the accuracy with which the listening panel guessed the weight of any one speaker was very low and probably did not exceed the accuracy expected from pure guessing. Lass et al. (1980) claimed again to have found evidence for accurate estimations of speaker weight, but a reanalysis of Lass's data by van Dommellen (1993) again suggested that this claim resulted from erroneous analysis, and that no correlations between estimated and actual weight were significant. However, interestingly, there were strong correlations between the judgments of weight made by different listeners. Thus, although listeners' judgments were not accurate, different listeners were consistent in their judgments, suggesting that they made use of the same principles to estimate size. This finding is consistent with the experimental results of this thesis, reported below.

In the animal kingdom, there is somewhat better evidence that F_0 provides an accurate cue to body size. Davies and Halliday (1978) reported a negative correlation between the body size and fundamental frequency of the croak of common toads, and found that this information influenced the outcomes of bouts of male / male aggression. Ryan (1988) reviews the evidence of a size / frequency relationship throughout the anuran amphibians (frogs and toads). However, even in anurans the

relationship between body size and fundamental may be more complex: females often have higher fundamental frequencies than males, although they typically outweigh them (e.g., McClelland & Wilczynski 1989, Walkowiak 1988, Zakon & Wilczynski 1988). In addition, some fish species appear to use fundamental frequency to assess strength (Ladich et al. 1992), although the sounds produced by fish are created not by a larynx but by vibrations of fins or the swim bladder.

Hauser (1993) found a significant negative relationship between body weight and fundamental frequency using data from 36 nonhuman primate species. Of course, this cross-specific correlation does not bear upon the question of whether or not F_0 provides an accurate cue to body size within any one of these species. In fact, in the seven species of the genus *Macaca* (the only genus with enough species to allow a quantitative test) Hauser examined, there was no relationship between body weight and F_0. Thus, although Hauser's data support the overall conclusion that larger animals make lower pitched sounds, the critical within-species question remains unanswered. Gouzoules & Gouzoules (1990) examined the relationship between frequency and body size in pigtail macaques (*Macaca nemestrina*), but their acoustic measurements appear to reflect neither F_0 or formants, leaving the interpretation of the negative relationship they found unclear in terms of meaningful acoustic attributes.

These data from animals suggest an intriguing hypothesis. Perhaps F_0 did provide an accurate cue for body size in our early mammalian ancestors. Given the importance of body size in animal social behavior (see e.g. Peters 1983, Schmidt-Nielsen 1984) and the pre-existing ability of the auditory system to estimate F_0, this provoked the development of a mechanism which used F_0 to estimate body size. Of course, once such a mechanism was in place, it paved the way for its own manipulation: given the relative ease with which larynx size and/or vocal fold length and mass can be modified independent of body size during the process of evolution, selection would favor vocalizers who devoted energy to enlarging the larynx rather than to enlarging the entire body. It seems quite plausible that this hypothesis accounts for the large difference in F_0 between human males and females: females are only 22% lighter than males, and only 6% shorter, but their average fundamental frequencies are almost double those of men in Lass's sample of American college students (Lass & Brown 1978). This difference is accounted for by the increased growth in larynx size and vocal fold length which occurs during male puberty (Negus 1949, Goldstein 1980). Apparently, there has been stronger selection for males to be large than for females (this could be due to the adaptive value of large body size in male-male aggression, or to female preference for large males, or both). Although one result of this pressure has been the actual size difference between human males and females, another has been the difference in <u>apparent</u> size conveyed by the lower-pitched male voice.

The modifications of larynx and vocal fold size in humans are quite modest compared to those of the large fruit-eating African bat *Hypsignathus monstrosus* (commonly know as the hammerhead bat). This species is one of the largest in Africa (and in the world), but its vocal apparatus is unparalleled in all of the mammalian order: the larynx of the male is so enlarged that it occupies most of the chest cavity, having displaced the heart and lungs into the abdomen (Schneider et al., 1967). The sound produced by the hammerhead bat, frequently termed a croak (Schneider et al., 1966) was described more colorfully by Kingdon (1974, p. 171) as being "guttural, explosive and blaring, somewhat like the burst from an electric alarm". Because the mating system of this species involves female choice at "lek" sites, where males congregate and vocalize, and where the top few males obtain virtually all matings (Bradbury 1977), sexual selection in this species is probably very powerful, which presumably explains the extreme development of the larynx in this species (which also, incidentally, has a variety of other sexually dimorphic vocal adaptations, including laryngeal air sacs and an enlarged nasal resonating cavity).

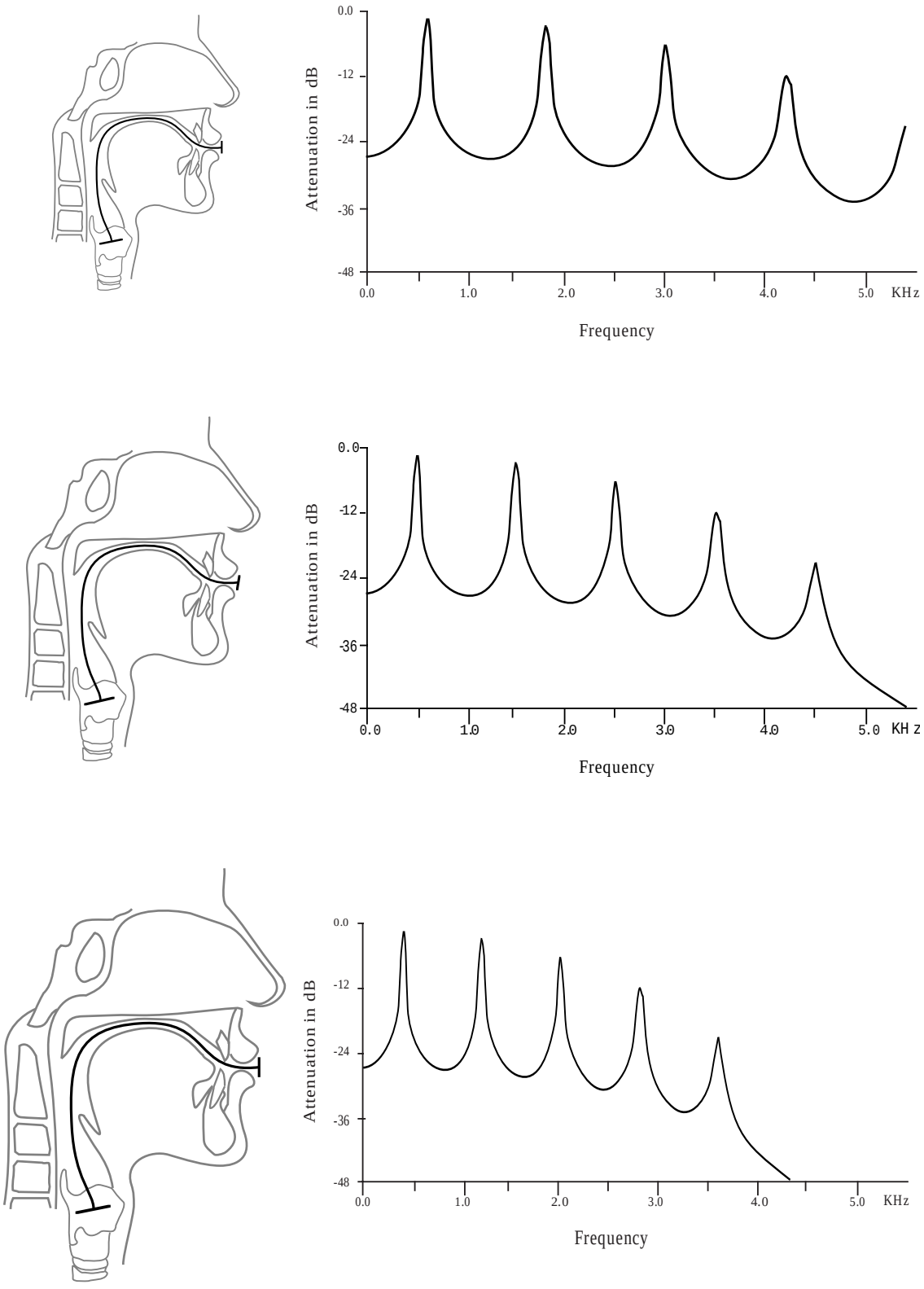


Fig. 5: Side view of the human vocal tract illustrating the effect of vocal tract length on the vocal tract transfer function

In summary, F_0 seems to serve as an accurate cue to body size in some frog species, and may play a role in the perceptual estimation of size in many more species, including humans. However, because larynx size can change independently of body size over evolutionary time, there is no strong relationship between body size and vocal fold length in many species. As a result, F_0 does not provide a valid cue to body size in these species.

Vocal Tract Cues

There are several a priori reasons to expect the length of the vocal tract to provide a robust and dependable cue to body size. Vocal tract length, which is (in mammals) the distance from the glottis to the outer portion of the lips, is strongly constrained by skull size, which is in turn allometrically coupled to total body size. Thus the actual length of the vocal tract will be tightly constrained by body size (although lip size and the position of the larynx are also important determinants). Because the length of the vocal tract controls (all other things being equal) the dispersion of formants in the vocal tract transfer function, formant dispersion should provide a readily-available acoustic cue to body size. Figure 5 illustrates the effect of vocal tract length on formant pattern.

Whether or not listeners make use of this cue is a different question. In humans at least we might expect that they do, since human speech depends crucially on formant locations (so we know that they are perceived), and the ability to understand different-sized speakers depends on an ability to adjust for the differences in formant dispersion which result. This adjustment process, known as vocal tract normalization, was first demonstrated by Ladefoged and Broadbent (1957), who showed that a probe word such as [bIt] would be perceived as *bit, bet, bat* or *but* depending on the formant dispersion of the preceding sentence "Please say what this word is..." (generated by a speech synthesizer). This experiment provided the classic illustration of the fact that listeners adjust their speech perception to the perceived vocal tract length of the speaker (see also Broadbent and Ladefoged 1960).

Because of the wide range in vocal tract lengths among human speakers (e.g., between a young child and a large adult male), the process of vocal tract normalization is absolutely essential in the perception of human speech. This is particularly true for children, since their shorter vocal tracts make it physically impossible for them to produce the same absolute formant frequencies as adults. For a child to imitate an adult vowel, it must therefore compensate perceptually for the vocal tract length difference between the adult word and their own. Given that there is presumably a well-developed perceptual mechanism which performs this task for the purposes of speech perception, it seems rather plausible that it would also be used for the purposes of body size estimation. This hypothesis is tested later in this chapter.

There is suggestive evidence that some frog species use the resonant frequency (or frequencies) of the supralaryngeal vocal tract to assess body size. Wilczynski and colleagues (1993) did a careful analysis of body size, larynx size and "dominant frequency" in the advertisement calls of three species of hylid frogs (genus *Hyla*). Dominant frequency was defined simply as the frequency of the highest-amplitude harmonic in the Fourier spectrum of the call; this frequency is known to be important behaviorally for this species. These calls are pulsatile, with low pulse rates (around 200 Hz), while the dominant frequency is typically a spectral peak between 3 and 6 kHz. Examinations of Fourier spectra and spectrographs of these calls strongly suggests that these calls would be well-described by a source/filter model, and that the "dominant frequency" is an indicator of a high-frequency resonance or formant which filters the low-frequency laryngeal source. Although dominant frequency did not correlate with larynx size, it showed a strong negative correlation with head size, suggesting that the size of the resonant chambers of the head determine dominant frequency as would be expected from the source/filter theory.

Another hylid frog species, the cricket frog *Acris creptians*, also has advertisement calls possessing a dominant frequency which appears to result from filtering by the supralaryngeal vocal tract.

Wilczynski et al. (1992) studied these frogs, showing that males exhibit a negative correlation between body size and call dominant frequency (females do not produce advertisement calls). Furthermore, the female auditory system is tuned to, and females prefer, calls with dominant frequencies slightly lower than the population mean. This suggests a female preference for large body size, as reflected by lower resonant frequency in the advertisement call, in this species. Unfortunately, despite the voluminous and detailed literature on anuran vocal communication, the potential significance of laryngeal vs. vocal tract cues for the evolution of communication in this group has not been carefully explored.

In birds, as mentioned above, the length of the vocal tract includes not just the size of the oral and pharyngeal cavities, but also the length of the entire trachea. Because the trachea is a flexible cartilaginous structure, this has provided an interesting possibility for vocal deception to birds which is not available to mammals: the trachea can be elongated and coiled up, thus decoupling tracheal length from total body size. This phenomenon of vocal tract elongation is commonly seen in Anatidae (ducks, swans and their allies), and was described in both American species of cranes as early as 1880 (Roberts 1880). In cranes, the trachea invaginates into the sternum (in most birds, the bones are hollow), which imposes at least some structural limit on vocal tract length. In contrast, in a passerine species *Manucodia keraudrenii* (the trumpet manucode), the trachea coils into the space between the skin and the outer chest walls, allowing it to undergo unrestricted elongation to an extreme not seen in any other species. These birds live in New Guinea, and are members of one of the most advanced bird taxa (birds of paradise). Most species of birds of paradise have ornate visual ornamentation; in comparison, the trumpet manucode is quite plain and crowlike. However, it appears that this tracheal elongation may qualify as an example of acoustic ornamentation. In both the trumpet manucode, and all of the other species which have been investigated, the elongation of the vocal tract is sexually dimorphic, with males having much longer tracts.

Summary of Possible Acoustic Cues to Body Size

Thus, there are a number of possible acoustic cues to body size, few of which have been experimentally examined. Lung or tracheal volume may provide a cue, either by setting the maximum duration of calls, or by creating tracheal formants which provide a cue to tracheal length. (In birds, the syringeal source is at the base of the trachea, and thus tracheal length is equivalent to vocal tract length). To the extent that vocal cord length and mass are correlated with body size, the lowest fundamental frequency of phonation an individual can produce could provide another cue to body size. This appears to be the case for (at least) some non-human species, but there is little empirical support for this association in humans. Finally, the length of the supralaryngeal vocal tract and its corresponding cue of formant dispersion seems on theoretical grounds to be a good candidate for an acoustic cue to body size.

In the experiment reported below, I tested the extent to which two acoustic cues, fundamental frequency and formant dispersion, influence human listeners' judgments of speaker body size. On the basis of earlier reports (e.g., van Dommellen 1993), it seemed likely that fundamental frequency would play a role in the perception of body size, despite the lack of evidence for a dependable relationship between F_0 and body size. Although no data have yet been published specifically examining the role of formants in estimates of body size, the logic described above makes it seem likely that formant dispersion should also provide a cue for size. Thus, these two variables were chosen as independent variables in the study which follows.

Experiment 1

Methods

Stimuli:

The stimuli for Experiment 1 were synthesized using the algorithms described in Klatt & Klatt (1990) using the Sensyn speech synthesis package on an Apple Macintosh Quadra 800. The sampling rate was 11.127 kHz quantized to 16 bits. Utterance duration was 500 ms. Two different mean fundamental frequencies were used (100 and 150 Hz). For increased naturalness, there was a slight downward frequency ramp over the first 100 ms, starting at 110% of the mean frequency (i.e. 110 Hz and 165 Hz, respectively). Random F_0 variation (termed "flutter" by Klatt & Klatt 1990) at 10% of the mean frequency was added to the F_0 contour. Other source parameters were left at the default values (50% open quotient, 200% rise/fall time quotient, 60 dB voicing amplitude, no aspiration).

The filter function had five formants in the locations predicted by a simple open tube model of the vocal tract for four vocal tract lengths (15-18 cm) (Lieberman and Blumstein 1988). Such a model approximates the shape of the unperturbed vocal tract, and corresponds perceptually to the schwa vowel. The vocal tract is modeled as a simple tube with a pressure maximum at the glottis and a pressure minimum at the lips. Sounds with wavelengths which match these constraints pass through the tube easily, while other wavelengths are damped (the energy being absorbed as heat by the tube walls). The lowest resonance predicted by this formulation is at a wavelength four times the tube length, corresponding to the first formant. Other resonances occur at odd integer multiples of this value (see Equation 3, above). The calculated formants for each vocal tract length (measured from the glottis to the lips) are given in Table 1. The bandwidth for any particular formant was held constant for all vocal tract lengths and are based on the TBFDA values for the 17 cm vocal tract; they are given at the bottom of Table 1.

Table 1: Formant values (in Hz) for Stimuli in Experiment 1

Length (cm)	First	Second	Third	Fourth	Fifth
18	465	1396	2326	3257	4188
17	493	1478	2463	3449	4434
16	523	1570	2617	3664	4711
15	558	1675	2792	3908	5025
Bandwidth:	60	90	150	200	200

The vocal tract lengths chosen span the normal range for adult males (Lieberman and Blumstein 1988). The filtering was performed by the cascade branch of the Klatt synthesizer with the number of formants set to five.

The quality of these stimuli was not particularly natural; no great pains were taken to make either the source or the filter exactly model human vocal output. As a result, the stimuli sounded somewhat harsh and mechanical, and were not readily mistaken for actual human vocalizations.

In summary, the stimuli consisted of eight 500 ms vowel sounds which had either a high or low fundamental frequency (100 or 150 Hz), and formant frequencies corresponding to one of four vocal tract lengths (15, 16, 17 or 18 cm).

Subjects

Subjects were 11 Brown University undergraduates with no medically-diagnosed hearing problems. They were paid for their participation.

Procedure

After synthesis the stimuli were transferred onto a digital audio playback tape using a Panasonic SV-250 portable DAT recorder via the analog outputs of an Audiomedia II D/A board. Each of the eight stimuli was presented six times, resulting in a total of 48 trials (presented in random order), with an ISI of 3.5 s. Playback to subjects was accomplished using the SV-250 DAT player to drive two Bose Roommate II powered speakers (identical signal from each speaker) to 1-4 subjects at a time. Subjects were seated comfortably 1-2 m from the speakers in a 6 x 8 m room with walls of cinder block and wooden baffles and a ceiling made of acoustic tiling. Subjects were asked to rate the apparent body size of the person producing the sound on the tape by circling a number between one and seven on a worksheet handed out before the experiment. The following instructions were read by the experimenter at the start of each session:

> This is an experiment to find out how we judge body size from the sound of people's voices. I am going to play a tape with the voices of several different individuals on it, each of whom will make several different sounds, and I want you to rate how big you think the person was who made the sound. You'll record your rating by circling the number on the worksheet, where "1" means a small person and "7" means a large person. I would like you to use normal sized people as the basis for these ratings, so "1" does not mean a midget and "7" a giant. There is no trick involved here, I just want you to say how large the speaker seems to you at a gut or intuitive level. If you don't know, just guess. Are there any questions?

The verbal instructions thus did not fix explicit end points, so each subject was free to use whatever cues he or she found salient to rate the speaker's body size on a scale of his or her choosing. At the end of the testing session I interviewed the subjects, asking them to volunteer which acoustic variables they thought they were using to determine speaker size.

Results

While reporting results I use the terms "formants" and "vocal tract length" as a shorthand for the manipulation of the length of the vocal tract of the computer-synthesized speaker, which resulted in changes in both the absolute frequency of each formant and the spacing between the different formants. "Vocal tract length" is the numerically-specified independent variable, which controls "formants", the acoustic feature perceptually available to the subjects. It should be kept in mind that this manipulation did not involve changing the overall pattern of the formant frequencies (which would change the vowel quality), but instead "stretched" or "compressed" the same evenly-spaced pattern. All sounds in this experiment retained the sound of the schwa vowel.

Subjects found the task to be quite straightforward: no subjects complained that it was difficult or impossible to know how large someone is by the sound of their voice. Subjects varied in how much of the seven-point rating scale they used, from a minimum of three points (N=1) to the maximum of all seven points (N=4) (mean range was 5.7 points). Subjects centered their responses, as expected, near 4 (the center of the rating scale) (mean = 3.8, s.d. = 0.68). Fig. 6 shows the frequency distribution of responses for all subjects combined. The ratings are approximately normally distributed so standard parametric statistics will be used to analyze the data. As is commonly the case, subjects tended to avoid the extremes of the rating scale.

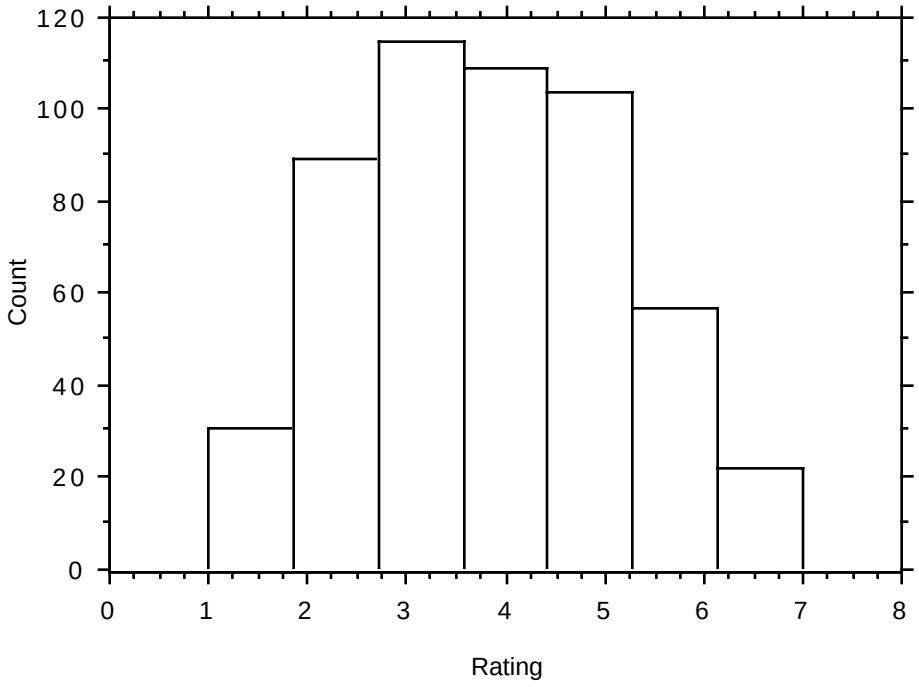


Figure 6: Frequency Distribution of Responses Combining all Subjects

To search for an effect of formants, fundamental frequency or stimulus repetition on body size ratings, I performed a 3-way repeated-measures ANOVA (the probability values given are adjusted using the Greenhouse-Geisser epsilon, which results in a conservative estimate of *p*). All ANOVAs were performed using the raw body size ratings as the dependent measure. The results, summarized in Table 2 and illustrated in Figures 7 and 8, show that there was a strong effect of both formants and F_0 on body size ratings.

Because each sound was presented six times, it was possible that there would be a learning effect (e.g., that subjects would learn to focus on, or ignore, one of the acoustic variables as the experiment progressed). I included stimulus repetition (which of the six presentations of a particular sound) in the ANOVA in order to test for such a pattern. The analysis revealed no effect of stimulus repetition. There were no significant interactions.

Regression analysis of vocal tract length vs. mean body size ratings (over all subjects) revealed a significant correlation between formants and size judgments for both the high and low F_0 stimuli (N= 24, Low F_0: R^2 = .561, F = 30.4, $p < .0001$; High F_0: R^2 = .597, F = 35.1, $p < .0001$). The slopes of these regression lines for the two sets of stimuli were nearly identical (see Fig. 7). Comparisons of the slopes of the regression lines of high and low F_0 for each subject individually showed that the slight difference in slopes between these sets of stimuli was not statistically significant (paired t-test: t(10) = .477, p = .6436). This was true even if only the slopes from regressions which were statistically significant were compared (N=14 significant regressions, 7 per grouping, unpaired t-test: t(12) = .625, p = .5436). Thus, there was a clear effect of formant frequency completely independent of the effect of fundamental frequency.

Table 2: Results of 3-Way Repeated Measures ANOVA for Expt. 1.
Dependent Variable: Body Size Rating (N = 11 Subjects)

Source	df	Sum Sqs	Mean Sq	F	p
Subject	9	218.333	24.259		
VT Length	3	77.817	25.939	28.566	.0001
Fundamental	1	456.300	456.300	63.967	.0001
Repetition	5	7.017	1.403	1.231	.3169
VT Length * Fundamental	3	6.517	2.172	1.978	.1734
VT Length * Stim Order	15	10.533	.702	.885	.4713
Fundamental * Stim Order	5	3.050	.610	.723	.5330
VT Length * Fundamental * Stim Order	15	20.233	1.349	2.332	.0646

Figure 7: Change in Mean Body Size Rating with Vocal Tract Length

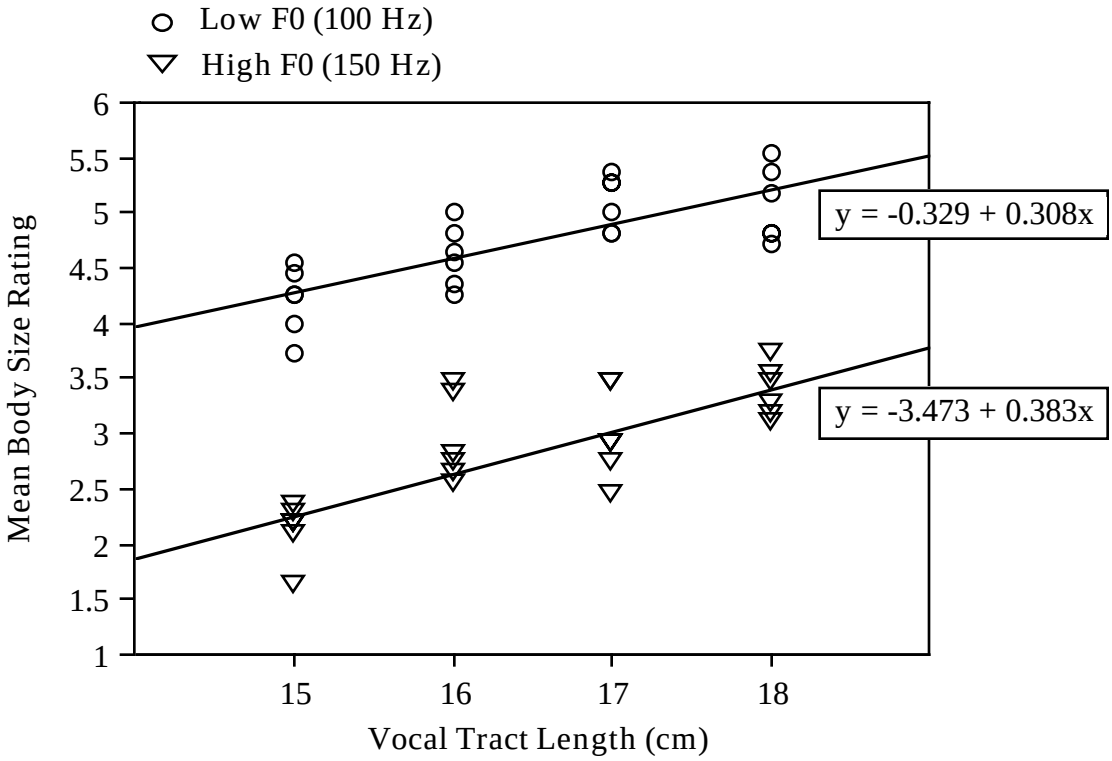
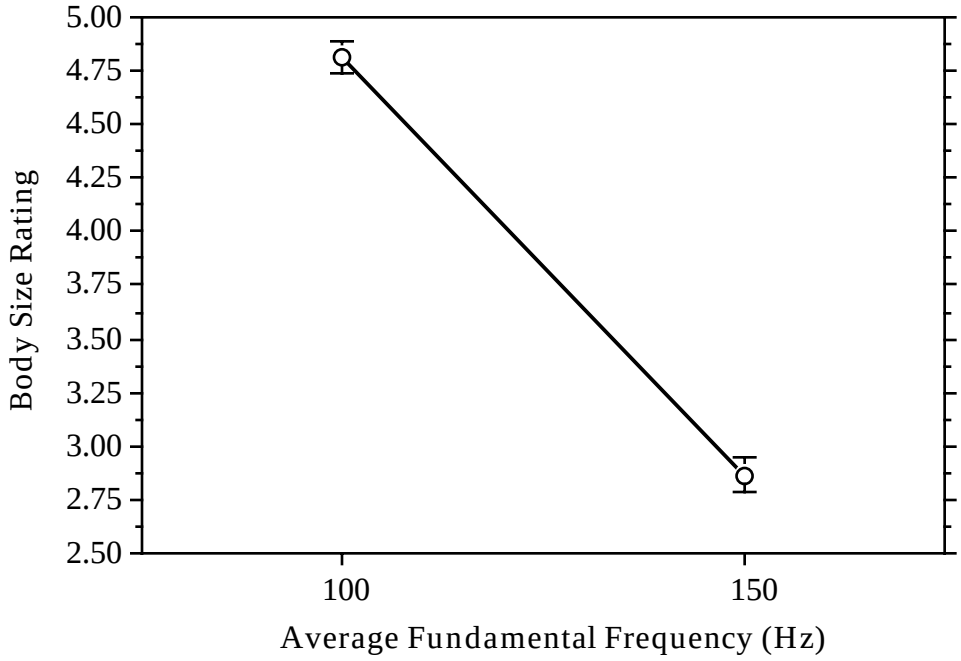


Figure 8: Change in Body Size Rating with F_0 (Bars show Std. Error)



As expected, subjects associated lower-pitched voices with larger body sizes (given the relatively large difference in fundamental frequency used here: 100 Hz vs. 150 Hz). The change in body size ratings that resulted was about 2 rating units, while the difference resulting from the formant manipulation (from 15 to 18 cm) was about 1 rating unit.

Body size ratings increased nearly linearly with synthesized vocal tract length, as can be seen in Figure 7. To determine whether there were any significant higher-order effects, I performed a trend analysis using the method of orthogonal polynomials. While the linear trend was highly significant ($F(1) = 8.06$, $p = .0068$), adding quadratic or cubic terms did not significantly increase the goodness of fit ($F(1) < 0.6$, $p > .45$). Thus, as predicted, the relationship between formants and body size rating is best described by a simple, linear function. This effect was very consistent across subjects: regression analysis for each individual showed that all subjects had a positively-sloped relationship between vocal tract length and body size rating (mean slope = .36, min = .19, max = .52).

To determine the extent to which subjects were aware of the acoustic bases of their body size ratings, I asked subjects which acoustic variables they thought they were using to determine speaker size. While three subjects volunteered that they used “timbre” or “resonance” to aid their estimates of body size, five subjects said they used only “pitch”, even after I suggested that they might be using some other cue like “sound quality”, “timbre” or “resonance” (the remaining three subjects were agnostic, saying they simply went by “how big the speaker sounded”). However, these subjective impressions seemed to bear no relation to the actual performance of the subjects: when I performed the same 3-way repeated-measures ANOVA on only the five subjects who claimed to be using pitch alone, they still showed a strong effect for formants ($F(3) = 9.81$, $p = .0015$). While it is possible that these subjects were using the term “pitch” to somehow include the timbral change caused by the formant manipulation, it seems equally likely that subjects are simply unaware of the voice quality cue at a conscious level.

Discussion

The results of this experiment indicate that both fundamental frequency of phonation and formant dispersion play a significant role in subjects' estimations of human body size. This is the first experiment of which I am aware in which listeners were asked to make judgments of body size based on computer-synthesized speech. The simple fact that they can easily and consistently perform the task is thus of considerable importance. Because most studies which have asked subjects to characterize the speaker have been based on real speech, in which there are a huge number of uncontrolled and unmeasurable acoustic variables, the conclusions that can be drawn from such research are at best tentative.

For example, Lass and his colleagues (1980) attempted to use low- and high-pass filtering to isolate "fundamental frequency" and "formants" from natural speech. Such experiments are flawed because high-pass filtered speech still contains extensive evidence about fundamental frequency, because high harmonics remain in the signal which provide a good cue to F_0 (we derive fundamental frequency from such cues all the time when speaking on the telephone or listening to low-quality transistor radios). This is known as perception of the “missing fundamental” (Schouten 1940). Similarly, most low-pass filtered speech still contains information about the lower formants (particularly for front vowels) due to the fact that filters have a pass-band of appreciable width. Experiments with natural speech are important, but they demand the use of more sophisticated signal-processing techniques (such as analysis and resynthesis with LPC) than have been deployed to date. The method used in this experiment allows us to avoid the "uncontrolled parameter" problem entirely: since we created the signals, we know the precise details of their acoustics.

The finding that body size ratings are negatively correlated with fundamental frequency is unsurprising given the intuitive sense most people have that larger people have lower voices. However, given that none of the studies which have actually measured body weight or height and F_0 have found a significant correlation, this finding might seem puzzling (because all of the speech samples used in this study were in the male range, and were perceived as male, the possible factor of male/female differences is ruled out). There are at least two approaches to reconciling these data. The first idea is that our intuitive notion is correct, and that the studies addressing the issue thus far have been flawed. The most obvious flaw in previous studies (i.e., Lass et al. 1978, 1980, Lass & Brown 1978, Künzel 1989, van Dommellen 1993) is that they allowed their speakers to use an F_0 of their own choice. Given that some speakers may have been more or less nervous under the recording conditions, and that anxiety and excitement are known to result in an increase in voice pitch (Williams and Stevens 1969, Scherer 1981), it is possible that real correlations between F_0 and size were obscured by differences in affect or other factors. A more satisfactory technique would be to ask speakers to produce the lowest and highest notes possible (though it is possible that differences in musical training between subjects could play a confounding role in this case).

A more radical suggestion is that F_0 and body size are truly uncorrelated in human beings, but that we retain an ancient perceptual mechanism (which we share with other species as primitive as frogs and toads) which (erroneously) makes this association. Of course, the fact that such a mechanism might be erroneous for body size judgments in our own species might not render it completely invalid, since it would still be useful in judging the body size of at least some other taxa (e.g., perhaps, dogs). Furthermore, the use of F_0 in body size judgments still provides some information if it is used across genders, due to the average body size difference between male and females in most cultures. In any case, it is clear that more research will be required before this issue is resolved, and the data above clearly indicate that human listeners indeed make use of fundamental frequency when estimating the body size of a speaker.

The positive correlation between body size judgments and formant dispersion demonstrated here provides the first demonstration in the literature that formants are used as a cue to a speaker’s size.

The relationship discovered is unsurprising from an acoustical viewpoint because formant dispersion provides an unambiguous cue to vocal tract length, which is in turn presumably correlated with total body size (see, e.g., Fant et al. 1972, Klatt & Klatt 1990). Given that the estimation of body size is presumably a more primitive function than speech perception, this finding raises the interesting possibility that the original use of formant perception was the estimation of body size. In particular, it suggests that a mechanism for estimating body size from formant dispersion may have provided a preadaptation for the process of vocal tract normalization, which is absolutely essential in speech perception and child language acquisition. Vocal tract normalization was supposed by Lieberman (1984) to be part of the unique human endowment for speech perception.

However, an alternative to the "size judgments as preadaptation" hypothesis above is that a recently-evolved, uniquely-human mechanism for vocal tract normalization gives body size judgments "for free", and that the influence of formants on size judgments documented here is thus an epiphenomenon. For the preadaptive hypothesis to have any weight, we need evidence that formant dispersion plays a role in the body size judgments made by other mammals. I offer such evidence in Chapter 3 of this thesis, and in the final chapter of this thesis I return to a detailed consideration of the evolutionary issues raised in the present discussion.

Chapter Summary

Acoustic theory allows us to predict that a number of aspects of the speech signal, or more generally of vertebrate vocalizations, could be correlated with the body size of the producer. These potential cues to body size include the duration and loudness of a vocalization (pulmonary cues), the frequencies of sub-glottal formants (a pulmonary cue), the fundamental frequency of phonation (in humans, a laryngeal cue), or the supra-glottal formant dispersion (a vocal tract cue). I examined the role of F_0 and formant dispersion in the judgment of body size by human listeners by creating computer synthesized speech sounds with different pitches and vocal tract lengths and asking subjects to rate the size of these "speakers". Both fundamental frequency and formant dispersion played a role in body size judgments; F_0 was negatively correlated with body size ratings (lower pitch meant larger speaker), while formant dispersion was positively correlated (longer vocal tracts meant larger speakers). The latter finding is particularly interesting in light of the importance of formants in speech perception, suggesting that formant perception may have initially evolved to aid body size judgments and then been "co-opted" for use in speech perception (the "preadaptation" hypothesis). However, it is also possible that the use of formant dispersion in body size judgments is a spin-off of a recently-evolved uniquely-human mechanism for vocal tract normalization. Experiments demonstrating that formant dispersion functions as a cue to body size in non-humans are necessary to exclude the latter possibility.

Chapter 2:

Vocal Tract Length and Phonetic Symbolism: the Arbitrariness of the Sign Reexamined

Plato's dialogue *Cratylus* opens with Hermogenes asking Socrates if the sounds of words are simply arbitrary conventions (as Hermogenes believes) or instead reflective in some way of their meaning (as held by Cratylus). Socrates argues forcefully for the latter, holding that although many words have arbitrary relations to their meanings, good words are distinguished by a correspondence between sound and sense; they have a sound which suits their meaning. To the contemporary reader, Plato's conclusion seems almost quaint: following Ferdinand de Saussure (1916), it is widely believed that the connection between sound and meaning in language is completely arbitrary, except for certain aberrant phenomena such as onomatopoeia. However, even at the time of publication of Saussure's "Course in General Linguistics", he was criticized for this stand by his contemporary, the eminent linguist Otto Jespersen, who pointed out that sound / meaning correspondences more subtle than onomatopoeia may play an important role in language, citing instances of non-arbitrary pairings between sound and meaning in core items of the vocabularies of English and several other Indo-European languages (Jespersen 1922, 1933).

The idea that units of speech smaller than words convey meaning (independent of their configuration in whole words), historically termed "phonetic symbolism" or simply "sound symbolism", has in this century received the attention of such luminaries as Roman Jakobson (1978, 1979) and Edward Sapir (1929), and has generated a substantial and diverse literature exploring phonetic symbolism in many languages from a variety of descriptive and experimental perspectives. Although this literature is replete with rigorous and convincing demonstrations of phonetic symbolism, the study of the phenomenon is conspicuously absent from contemporary mainstream linguistics, anthropology or semiotics, and a student attending an introductory course in any of these disciplines is unlikely to even encounter the idea. The main reason for the relegation of phonetic symbolism to the backwaters of language study, I believe, is that no one has yet offered a convincing explanation as to its underlying basis. While it is certainly interesting that a large random sample of the world's languages use high front vowels (e.g., [i]) in words for "this" and use back vowels ([u] or [a]) in words for "that" (Woodworth 1991), the question of why such a pattern should exist remains. Various mechanisms have been postulated to explain different types of phonetic symbolism, ranging from a postulated similarity between the vocal gesture and the action denoted, to vague notions of synesthesia or "frequency coding".

In this chapter I consider one type of sound symbolism: the association between vowels and size. This is easily the most commonly-cited example of phonetic symbolism, starting with Jespersen's (1922, 1933) observation of the overwhelming association of the vowel [i] with objects of small size and diminutives, and the use of [u] or [a] to denote larger size. After reviewing substantial evidence that this pattern is real, not only in English but in many other languages as well, I propose that the vocal tract length differences in the production of different vowels leads to an impression of greater or lesser size due to the association made by listeners between vocal tract length and body size. I then present experiments which support this hypothesis and rule out a number of competing explanations. I conclude by challenging the notion of the arbitrariness of the sign and proposing that phonetic symbolism plays an important role in human language.

Terminology

A number of terms of widely varying utility have been developed in connection with sound-symbolic concepts. It seems useful to define two here. The "phonestheme" is the basic unit of associated sound and meaning in sound symbolism (Householder 1946), and may consist of a single phoneme or a group of phonemes. For example, /fl-/ in word-initial position was held by Jespersen (1922: p 400) to be a phonestheme expressive of movement (e.g., flow, flake, flutter, flap, flicker, fling, flit). In general, the phonestheme represents a level of linguistic organization intermediate between that of the phoneme and the morpheme (if "fl-" were considered a morpheme, we would correspondingly have to

posit the "forlorn would-be morphemes" "-ow", "-ake", "-utter", etc. to complete the words in the list above (Ladd 1978, Markel & Hamp 1960)). Thus a phonestheme is the smallest unit with a specific meaning in language; in the case of vowel symbolism, a phonestheme can consist of a single phoneme. (Other proposed terms for this concept include "psychomorph" (Markel and Hamp 1960) and "submorphemic differential" (Bolinger 1965); the term was coined, as "phonæstheme", by Firth (1930).

An "ideophone" is a word making use of sound symbolism (Doke 1927). The term was originally proposed by Doke to describe words in the Zulu language, in which ideophones function as a distinct morphological class. The term was widely adopted for other African languages, and later for other language groups such as Korean, but has not yet been widely applied in Indo-European linguistics. (See Fordyce 1988 for a more detailed description of the evolution of the term "ideophone"; Westcott (1987) proposed the similar term "holostheme" for words made up wholly of phonesthemes).

There is one further distinction which has rarely been made (though see Fordyce 1988 and Brown 1958: p. 130) which seems capable of clarifying debate on the issue. I propose to distinguish between "natural" (or "mimetic") sound symbolism, which has its roots in a physical correspondence between sound and meaning (the most extreme version being represented by onomatopoeia), and "conventional" sound symbolism, which results from a human tendency to impute meaning to sounds (Bolinger 1950). As a result of this tendency, during the historical development of a language, listeners infer sound-symbolic structure in the lexicon, leading to the coalescence of clusters of similar-sounding words with similar meanings. I expect mimetic sound-symbolic pairings to be universal. In contrast, although the tendency which leads to conventional sound symbolism is probably universal, the specific sound/meaning pairings which result will be idiosyncratic and unique to a particular language. Much of the confusion and apparently negative evidence regarding sound symbolism in the literature is a result of a failure to make this distinction.

Fordyce (1988) proposes the terms "transparent iconicity" for onomatopoeia (which seems unnecessary given the preexisting term), "translucent iconicity" for mimetic sound symbolism, and "opaque iconicity" for conventional sound symbolism, based on terms developed by Bellugi and Klima (1978) to describe iconicity in sign language. However, conventional phonetic symbolism is not iconic at all (since there is no diagrammatic mapping between sound and meaning), and in the interest of keeping the discussion of sound symbolism as transparent as possible, I avoid using these terms here.

Natural vs. Conventional Phonetic Symbolism

Natural or mimetic sound symbolism is most prototypically displayed in onomatopoeia, translated from Greek literally as "the making of a name" and defined in Webster's as "the formation of a word by imitating the natural sound associated with the object or action involved; echoism". Given that the vocal apparatus is extremely flexible and capable of producing a wide variety of sounds, both tonal and noisy, it is unsurprising that we are able to imitate a wide variety of sounds from the natural environment. However, onomatopoeia does not involve slavish imitation of a sound using all of the available acoustic resources of the vocal tract, but instead maps the natural sound into the phonological system of the speaker's language. For example, the two-syllable song of the phoebe (*Sayornis phoebe*) can be imitated by whistling, but this is not an example of onomatopoeia. The common name of the bird is, because it is a phonologically normal word, and the stress on the first syllable mimics the higher pitch of the first note of the bird's song. Words formed by onomatopoeia are typically assumed to follow the phonological rules of the target language, but little research has been done to test this idea. Bladon (1977) studied spontaneous onomatopoeia by playing sounds and asking speakers of five different languages to describe them verbally. Although most of the words thus obtained obeyed the phonological rules of the speakers' languages, there were a few examples which did not. Similarly, Orr (1944) and Annamalai (1968) noted the resistance of onomatopoetic words to sound change, which could lead to phonological irregularities for these terms in some cases.

Onomatopoeia is widespread in the world's languages and its existence is unquestioned even by such linguists as Saussure (1916: p. 69). However, it has a rather limited domain of applicability, being definitionally restricted to use with objects or actions which produce sound. This should not lead us to underestimate its importance, however. For example, onomatopoeia exists in a very highly developed form in Japanese and Korean, where thousands of onomatopoetic words are attested. (The only Japanese writer to receive a Nobel Prize in literature, Yasunari Kawabata, is famous for his elegant and expressive use of onomatopoeia.) The extent to which onomatopoeia follows general cross-linguistic rules for mapping natural sounds into speech sounds has not been thoroughly investigated. Bladon's 1977 study suggests that the speakers of English, German, French, Turkish and Japanese use the same general rules to name sounds, and he explicitly formulates these rules in a testable form. Unfortunately, even a casual examination reveals much cross-linguistic variability in onomatopoetic terms (e.g., the sound of a cock crowing is "cockadoodledoo" in English and "kikeriki" in German), suggesting that although there are rules underlying onomatopoeia which structure form, they are less than 100% predictive.

Thus the concept of onomatopoeia is widely accepted, but its restriction to objects or actions which create sound leads most language researchers to believe that it plays a minor and perhaps deviant role in language in general. The idea that phonetic symbolism can operate along non-acoustic dimensions of meaning is much more controversial. Henceforth, I consider the terms phonetic symbolism (= sound symbolism) to exclude onomatopoeia as defined above (as does French 1977). I thus use "natural" or "mimetic" sound symbolism to refer to a correspondence between a speech sound and a non-acoustic dimension of meaning, such as shape, color, level of activity or (in the case of my experiments) size. For a particular sound / meaning pair to be held as an example of mimetic sound symbolism, the basis of the pairing must reside in a clearly-defined and relatively straightforward correspondence between the physical nature of the sign and the signified, as transduced by perceptual mechanisms amenable to psychophysical investigation. By this standard, the best candidate for mimetic symbolism is vowel symbolism for size, which I demonstrate later in this chapter. Because examples of mimetic symbolism are rooted in physics and general psychoacoustic mechanisms, we expect them to be general features of human communication, and latent even in those languages in which they do not happen to be expressed.

In contrast, conventional phonetic symbolism derives from a hypothesized inherent tendency of speakers and listeners to <u>impute</u> meaning to speech sounds, assuming that words which sound similar have similar meanings and thereby introducing regularity and order into the lexicon. For example, given that a large number of words in English which begin with "sl–" have a negative connotation (e.g., slime, slug, slobber, slur, slouch, slum, slovenly, slut), a listener exposed to a new "sl-" word will assume that it too has a negative connotation (probably influencing such recent coinages as "sleaze"). Given that many English verbs ending in "-ash" have a denotation of violent movement (crash, dash, smash, clash, bash, trash, mash), a listener can assume a new verb with the same ending will have a similar meaning (e.g., the recent appearance of the transitive verb "to trash" meaning to vigorously attack and demolish: to trash a house or trash someone's reputation). Thus, a few nearly synonymous words which sound similar can be enough to cause other words to undergo shifts in meaning, and can influence the coinage of new words; during the historical development of a language a small, randomly-planted seed can lead to the formation of a more extensive constellation of sound symbolic terms. As Bolinger (1950:p 128) put it, such terms "exercise a kind of magnetic attraction one upon the other". Such a process can lead in time to a significant number of sound symbolic terms in a language, which could play an important role in semantics, lexical access, and creative uses of language such as poetry and word coinage. However, we would not expect these patterns of conventional sound symbolism to hold across different languages, but to be unique and idiosyncratic to the language under study.

Conventional sound symbolism encompasses a broad range of possible levels of connection between the sound shape of a word and various aspects of its meaning (both connotative and denotative). Although I focus in the remainder of this chapter on the connections between single phonemes (vowels) and denotative meaning (size), there are many possible connections between larger units of sound structure and meaning which still deserve the rubric "sound symbolism". Examples include the frequent

use of syllable reduplication to signal diminutives, hypocoristics or other expressives, the pattern of /CeCo/ which signifies "foolish" in a host of Spanish words (Malkiel 1990), or even the use of stress to signal the distinction between nouns and verbs in English (Kelly 1992, Liberman & Prince 1977, Sherman 1975). Malkiel (1990) proposes the term "secondary phonosymbolism" for this phenomenon, though it might be more informatively termed "structural" sound symbolism.

Sound symbolism may also act as a force against sound change: the Latin *cacare*, for example, has changed to *chier* in French through a process of regular sound change, but the nursery noun *caca* has not changed since Roman times (Orr 1944).

It may be necessary to point out that the type of sound symbolism I am calling "conventional" is still a far cry from Saussure's concept of the arbitrary or conventional sign. In Saussure's example "the idea of 'sister' is not linked by any inner relationship to the succession of sounds s-ö-r which serves as its signifier in French...it could be represented equally well with any other sequence" (1916: p. 69). In words exhibiting conventional phonetic symbolism, this is not true: the sequence of sounds is related in a precise way to the word's meaning. Although the exact form of this linkage is arbitrary, within the context of a given language the relationship is lawful and non-arbitrary.

A Research Strategy for Sound Symbolism

The two types of sound symbolism distinguished above, natural and conventional, are quite different and their investigation demands different experimental and descriptive techniques. Much of the research in sound symbolism suffers from a kind of "grab-bag" approach: instead of focusing on particular examples of sound symbolism, experimenters have studied the idea of phonetic symbolism with long lists of arbitrarily chosen words. Here I propose a different research strategy. Initially, we hypothesize that a supposed example of sound symbolism (e.g., that the [i] vowel is associated with small size or short distance, or that "-ash" connotes violence) is of the conventional sort, and thus unique to the language in question. To determine if it is a *bona fide* example of sound symbolism of either sort, we have at least three types of tests available. We can perform a distributional analysis of words exhibiting the sound symbolic pattern in the language in question (does the number of words exhibiting the pattern greatly exceed the number expected by chance?). We can examine the growth or change in vocabulary exhibiting the pattern using the techniques of historical linguistics (have the meanings of similar-sounding words grown closer? Have words with overlapping meanings changed their sounds to suit?) Finally, we can create new words which exhibit the pattern, and ask naive subjects to describe their meanings. In the best case, all three techniques would be used to verify the existence of a postulated sound-symbolic pattern, but some languages may be inadequately documented for the application of the historical approach.

If these lines of investigation support the hypothesis of conventional sound symbolism, we may next attempt to construct a second hypothesis, postulating that the observed sound symbolism is an example of mimesis. One finding which would encourage such an endeavor would be a demonstration that speakers of unrelated languages are able to guess the meanings of these sound symbolic words at a level better than chance. To demonstrate mimetic sound symbolism, the researcher should first specify a connection between a specific acoustic characteristic and the meaning of the sound symbol, then use a speech synthesizer to create controlled stimuli which either exhibit this acoustic trait or do not. Using stimuli synthesized by computer is necessary to avoid the criticism leveled at many phonetic symbolism experiments since Sapir's (1929) study: that non-phonetic cues such as tone of voice or pitch could be used by subjects to guess word meanings. Following stimulus construction, standard experimental techniques can be used to test for the postulated sound/meaning connection by playing the stimuli to naive subjects to test for the presence or absence of the predicted meaning. If possible, the task should be constructed in the manner of a psychophysical experiment, asking subjects to rate the stimuli on some scale (rather than simply asking them to verbally record the meaning of the stimuli, which would encourage the operation of conventional sound symbolism). If the stimuli exhibiting the acoustic pattern show a sound symbolic effect, and the stimuli lacking it do not, we can conclude that there is a

psychophysical basis for the observed sound symbolism. It will then be interesting to examine other, unrelated languages to determine how widespread the pattern is.

In the remainder of this chapter, I apply the research strategy outlined above to vowel symbolism for size. I first review a number of studies which document that the vowel/size connection exists as an example of (at least) conventional sound symbolism in English. I then suggest that it is an example of mimetic sound symbolism, proposing that vocal tract length differences between the vowels explain the sound/meaning connection. I describe three experiments which support this hypothesis, and conclusively refute several previous proposals. Finally, I describe a comparative linguistic study which shows that a random assortment of the world's languages show the same pattern of vowel/size symbolism.

Vowel/Size Symbolism

By far the most extensively-studied and widely-accepted example of phonetic symbolism is the use of vowels to connote size or distance, in particular the use of [i] or [I] to suggest small size and [u] or [o] for large size. This connection was first mentioned by Tolman (1887) in his "Laws of Tone-Color in the English Language"; but it was first developed in detail by Jespersen (1922, 1933) in a relatively informal but nonetheless compelling fashion: he reported an extensive set of words, in a variety of Indo-European languages, where [i] was associated with small size, weakness or insignificance. In addition, Jespersen (1922) suggested that this vowel symbolism is particularly active in children, citing the interesting example of a German child who had concocted the term "lakeil" for "chair", and later used "likil" to refer to a small doll's chair, and called his grandfather's armchair "lukul". Similarly, the child's father "papa" was dubbed "pupu" when he donned a large fur coat. Orr (1944) used similar arguments and examples to reach the same conclusion.

Jespersen's observations were put on a firmer footing by the first reported experimental study of sound symbolism, performed by Sapir (1929). Sapir introduced a simple but powerful technique: he created new (nonsense) words, and asked listeners to rate the meanings of these words on some scale. Sapir use two variants of this technique: in the first he read words like "mal" and "mil" and asked subjects to indicate which referred to a larger object. In the second, he specified a nonsense word and then asked subjects to describe the meaning of various permutations of the word as he changed the vowels. For example, he specified that the term "mila" meant "brook". Then he asked subjects to specify what "mili" meant (most said a smaller, faster stream), or what "molo" meant (most said a large calm river or lake). Sapir found a striking agreement among his subjects (who included children of many ages and several native Chinese speakers in addition to adult English speakers) that [i] connoted smaller size, while [a] referred to larger objects.

Sapir's work was followed by that of his student Stanley Newman (1933), who used more rigorous statistical techniques and further experiments to extend Sapir's results. Newman studied both a wider range of vowels and some consonants. His analysis of the results allowed him to arrange the vowels along a continuum of implied size: starting with the smallest size, the continuum was ordered [i], [e], [a], [u], [o].

Several studies have examined size symbolism in the course of studies with a broader scope: sets of adjectives falling along the three axes (evaluation, activity and potency) of Osgood and Suci's (1955) "semantic differential". Since the "potency" dimension is congruent with size, all of these experiments test for size symbolism. Birch and Erickson (1958) created CVC nonsense words and had subjects rate them on scales of large/small, fast/slow and clean/dirty. Their results for the last dimension (evaluative) were negative, but the first two dimensions yielded statistically-significant consistency among subjects. The pattern of results (for size) was again in the order "i, e, a, u, o".

Johnson (1967) used a completely different technique to demonstrate vowel symbolism for size: he simply asked subjects to "write down all the words you can think of that denote, or suggest, smallness of

size on one side of the card; largeness on the other side of the card". Johnson's results fell on the same vocalic continuum as did Newman's results (described above).

The consistency of these studies is quite compelling, and the vowel/size continuum proposed by Newman now stands as the best-documented and least controversial example of phonetic symbolism currently available. It is, at least, a solid example of conventional sound symbolism. However, several studies suggest that vowel/size symbolism may be an example of mimetic sound symbolism as well.

Huang, Pratoomraj and Johnson (1969) used the same technique as Johnson (1967) (asking subjects to list all the words denoting or suggesting smallness) to document similar patterns of vowel/size symbolism in English, Chinese and Thai. Slobin (1968) found that adjectives chosen from Thai, "Kanarese" (= Kannada, a Dravidian language) and Yoruba were translated at above-chance levels by English speakers; magnitude symbolism was one of the semantic domains sampled. Klank et al. (1971) extended this translation technique, with the same results, to Czech, Hindi, Japanese and Tahitian.

Ultan (1984) surveyed 136 languages, including members of virtually all known language phyla (his sample was somewhat skewed towards Amerindian languages), searching for an association between particular phonemes and diminutive forms. Although many of the languages he examined showed no conclusive evidence of vowel/size symbolism, those that did virtually all exhibited the predicted pattern of high and/or front vowels for diminutive size (83% of the sample). Ultan's results suggest that a universal tendency exists for a pattern of vowel/size symbolism like that shown in English, although this tendency is not necessarily expressed in any particular language.

Further suggestive evidence comes from Woodworth's (1991) finding of a statistically-significant degree of vowel symbolism in deictics (words which indicate location relative to the speaker, e.g., "this", "that", "here" and "there" in English). Although this is, strictly speaking, not size symbolism, the dimensions of size and distance are similar enough to make suggestive her finding of a preponderance of high front vowels ([i] and [e]) in near deictics ("this") and back vowels ([a] and [u]) in far deictics ("that") (see also Ultan 1984 in this respect).

These comparative linguistic data indicate that vowel/size symbolism is consistent across a wide range of languages, suggesting that the pattern could be an example of mimetic sound symbolism. To test this idea, we must now construct a hypothesis which makes explicit what acoustic aspect of the vowel is responsible for the size symbolism.

Explanations for Vowel/Size Symbolism

A wide variety of explanations have been offered for vowel/size symbolism, starting with Sapir's (1929) suggestion that the small oral volume of [i] and larger oral volume of [u] or [a] are responsible for a kinesthetic impression of size associated with different vowels. Similarly, Sir Richard Paget suggested (1930: p. 137) that "the small front cavity, made by the tongue, to express smallness... naturally result[s] in the audible vowels i or I", thus accounting for vowel symbolism by "mouth pantomime". Sapir's student Newman proposed a post-hoc explanation based on a combination of three factors: the articulatory position of the tongue (front to back), the frequency of "vocalic resonance", and the size of the oral cavity. Orr (1944) suggested that the "greater muscular tension" or "lesser resonance" of [i] and [I] account for its association with small size, though he fails to make explicit why this should be so.

Vague notions of synesthesia form the basis for another class of explanations for vowel symbolism. A connection which seems to have particularly fascinated Roman Jakobson was that between colors and vowels, perhaps the prototypical example of synesthesia in speech (e.g., Reichard, Jakobson and Werth 1949, Jakobson 1978, Jakobson and Waugh 1979). Because "darkness" and "brightness" of vowels have been associated with large and small size, respectively, a chain of synesthesias has been held to explain the connection between vowels and size (see French 1977 for an example of this reasoning). This

proposition is exceedingly hard to test experimentally, there being no way I know of to manipulate the process of synesthesia.

In general, the explanations covered so far are so vague and makeshift as to make testing them impossible. Some more recent propositions are explicitly acoustic and thus lend themselves better to experimental testing. Several researchers (French 1977, Ohala 1983, Woodworth 1991) have proposed that a general "frequency code" explains several types of size symbolism: high frequency sounds connote small size, due to the fact that small animals make high frequency sounds and large animals make low frequency sounds. Although this hypothesis is impressively general (accounting for both tone symbolism, where high tone suggests small size in tonal languages, and vowel symbolism as well), it has several problems. The first is that there are difficulties with the "small animal = high pitch" rule. Fundamental frequency is correlated with the size of the vocal folds (in mammals) or syringeal membranes (in birds), which can vary independently of body size. In humans, there is no correlation between body size (measured by height, weight or surface area) and speaking fundamental frequency (Lass & Brown 1978, Cohen et al. 1980, Künzel 1989). However, an examination of the entire primate order showed the expected negative correlation between body size and fundamental, although the strength of correlation varied considerably among different taxonomic groupings (Hauser 1993). Even to the extent that a cross-specific correlation exists, it would always have exceptions. To cite just one example with common New England birds, the song of the cedar waxwing, *Bombycilla cedrorum* (weighing 33 g) is about 7 kHz, while the smaller white-throated sparrow *Zonotrichia albicollis* sings at less than 3 kHz but weighs just 26 g (Dunning 1993). Would a child exposed to exceptions such as these develop an inverted perceptual basis for sound symbolism?

In a variant of the frequency code hypothesis proposed by Ultan (1984), the correspondence between high frequency and small size results from the short wavelengths of high frequency sound. While the physical correlation between high frequency and short wavelength is indisputable, it is imperceptible without instrumentation: our percept of frequency is tied to the speed of vibration of the tympanum and not a measurement of wavelength. From a conscious point of view, the inverse correlation between frequency and wavelength is unknown to many humans who nonetheless are sensitive to sound symbolism (e.g., the nine-year old children in Sapir's (1929) study).

More problematic for all of these "frequency code" hypotheses is the conflation of fundamental frequency, a correlate of pitch, with formant frequencies, a measure of timbre. Although pitch can be specified as a single frequency value, the complex spectral patterns underlying vowel perception cannot: at least two formants are required to unambiguously specify most English vowels (and three to specify all of them). French (1977) seems to have missed this point; holding out the dubious fact that "[i] has one of the highest fundamental frequencies" as an explanation of vowel symbolism. Peterson and Barney's classic 1952 study of vowels shows this pattern only for women's voices; in men and children, [u] has the highest fundamental. In any case the differences in fundamental are quite small, and it is widely agreed that pitch is unimportant as a determinant of vowel quality (e.g., Fujisaki & Kawashima 1968, Ryalls & Lieberman 1982), so French's proposition lacks explanatory power. However, it is possible to devise a metric which collapses multiple formant values into one "frequency" (e.g., by subtracting F1 from F2, as suggested by Ohala 1983) or to restrict attention to a single formant (e.g., F2, as suggested by Woodworth 1991). Unfortunately, the "frequency code" theory leaves unspecified the grounds for these restrictions or transformations. However, both the "F2 alone" and the "F2 minus F1" theories are explicit and testable, and in the experiments below I shall contrast them with my hypothesis, which holds that the acoustic correlate of vocal tract length (formant dispersion) is the salient acoustic variable in vowel symbolism.

The production of vowels is dependent on the shape and length of the vocal tract between the glottis (which is the space between the vocal cords in the larynx) and the lips. The sound emitted from the larynx (the "source") is modified as it passes through the supra-laryngeal vocal tract (the "filter") to produce the final vowel sound, which radiates out from the lips (Fant 1960). The cross-sectional area function of the vocal tract (hereafter, its "shape") determines the pattern of formant frequencies, while

the length of the tract determines their absolute spacing along the frequency axis. The absolute volume of the vocal tract has little effect on the formant pattern (affecting mostly the bandwidths of the formants, which have little perceptual importance (Stevens and House 1955, Dunn 1961)); it is the relative volumes of different segments of the tract which matter. If the shape of the tract is held constant, increasing its length has a very specific effect on its acoustic transfer function: the lowest formant is shifted downward in frequency, and the spacing between higher formants decreases by a proportional amount. This effect of length is independent of the area function of the tract, and increasing the length simply "compresses" the spectral pattern, preserving its overall shape. Similarly, a decrease in vocal tract length leads to a "stretching" of the spectral profile. The important point is that spectral shape and overall formant spacing can be independently manipulated by varying the vocal tract shape or length, respectively.

Vowel production classically involves the manipulation of the vocal tract area function, moving the formant frequencies around relative to one another. However, vocal tract length also varies in an orderly way with different vowels. For example, while saying [u], the lips are rounded (making the opening between them small) and protruded (increasing the absolute length of the vocal tract). Both of these maneuvers have the same acoustic effect of lengthening the tract (Stevens and House 1955); and they accompany the production of [u] not just in English but in the vast majority of human languages (Maddieson 1984). In contrast, during the production of [i] the lips are relaxed or actively retracted in most of the world's languages, and this has the result of shortening the vocal tract. Lip protrusion is not the only mechanism by which tract length is manipulated: larynx lowering can also be used to increase the vocal tract length (Fant 1960, Stevens & House 1955). An additional type of vowel-specific length manipulation was described in Fant (1960: p. 114). When producing [u], the tongue body is pulled upward and backwards in the oral cavity. Air must flow around the tongue body, and thus must travel further from the larynx to reach the lips. The acoustic result is a lengthening of the vocal tract relative to more open vowels such as [a].

Human listeners use vocal tract length to estimate body size (see Chapter 1 of this thesis). Vocal tract length should provide a robust cue for body size because it is dependent (in mammals) on the size of the head and neck, which are closely tied by allometric constraints to the size of the rest of the body (unlike lengthening of the vocal cords, which is less constrained by allometry and requires a minimal energy input). I propose that vowel/size symbolism results from the variation in vocal tract length during the production of different vowels, which is interpreted by the psychophysical mechanism described in Chapter 1 as a variation in size. Thus, the [i] vowel is associated with small things because it is produced by a shortened vocal tract, which is perceptually correlated with smaller size. The vowel [u] seems large due to the longer vocal tract required to produce it. Because vocal tract length changes effect formant frequencies in a manner quite different from the competing hypotheses outlined above, we can devise a set of experiments to test between these hypotheses.

Introduction to the Experiments

The experiments described in this section constitute a test of a particular hypothesis: that vowel symbolism for size results from the fact that vocal tract length varies during the production of different vowels. Because vocal tract length influences subjects' estimations of body size, with shorter vocal tract lengths being associated with smaller size estimates (Chapter 1 of this thesis), the shorter tract length involved in the production of the vowel [i] should lead to a judgment of smaller size. I contrast this hypothesis with several alternative proposals from the literature: the "F2 alone" hypothesis and the "F2 minus F1" hypothesis (Woodworth 1991, Ohala 1984), and the idea that size symbolism derives solely from language convention (e.g., Taylor 1963) (summarized above).

To test between these hypotheses, I created two sets of vowel stimuli. In the first set, vocal tract length is held fixed, so that all vowels are produced with a vocal tract of the same length. If the vocal tract length associated with a vowel is responsible for its size connotation, this set of stimuli should yield no effect for vowel. Alternatively, if absolute height of F2, or the difference between F1 and F2,

is responsible for vowel/size symbolism, a significant difference for different vowels should be found. Experiment 2 tests these predictions. In the third experiment, vocal tract length was varied. Because the differences between the second set of stimuli are minuscule with respect to the "F2 only" or "F2 minus F1" hypotheses, these hypotheses predict little difference in experimental outcome. In contrast, the vocal tract length hypothesis predicts a significant effect for vowel in Experiment 3, unlike in Experiment 2.

Experiment 2 Methods

Stimuli:

A two-step process was used to create the stimuli used in this experiment, modeled on the source/filter theory of speech. First, a Klatt synthesizer (KLATTSYN88: Klatt & Klatt 1990) was used to synthesize two naturalistic glottal waveforms, with mean fundamental frequencies of 100 and 150 Hz. These source waveforms were then filtered using custom-written software to produce the final stimuli presented to subjects. This two-step process had the dual advantage of simplifying the preparation of the stimuli and ensuring that the fundamental frequency and vocal tract length factors in the experiment were indeed completely independent. Because Experiment 1 of this thesis showed that body size ratings increased approximately linearly with vocal tract length, only two extreme vocal tract lengths were used in this experiment (15 and 18 cm), corresponding to a small and a large male, respectively.

Vocal tract configurations corresponding to the following four vowels were used: [a], [i], [u] and a degenerate one-tube vowel approximating a schwa, [{]. The "schwa" was added to facilitate a comparison between these results and those of Experiment 1, but it is not based on a physiologically-realistic vocal tract shape or representative of a real [{], and thus the results for this vowel are not germane to the size symbolism hypothesis. The stimuli span the vowel space with the vowels [a], [i], and [u] forming the corners of the vowel triangle. Because these three vowels are the most common in the world's languages (of the 317 languages studied in Maddieson (1984), 274 had [a], 271 had [i], and 254 [u] - the next most common vowel was [o], in 133 languages), and appear in the vast majority of the worlds languages, this selection of vowels should provide results which generalize well to other languages.

I strove to make the stimuli sound as natural as possible (given the static formant values used: natural vowels virtually always have some change in formant frequencies over the course of the vowel), using as an example a natural [a] vowel spoken by me and adjusting the glottal source parameters by ear until the synthetic waveform sounded similar to the natural one. The specifications for the glottal source are thus considerably more complicated than in the previous experiment (all are given in terms of input parameters to KLATTSYN88). The vowel duration was 700 ms, with a sampling rate of 11,025 Hz (this was a slight increase from Expt. 1, in order to impose less attenuation on the high formants) and 16-bit quantization. All stimuli were amplitude-normalized so that each vowel had the same peak amplitude. Two different mean fundamental frequencies were used (100 and 150 Hz). Though the F_0 was mostly flat, there was a downramping from 105% of the mean F_0 over the first 75 ms, and a downramping to 95% of the mean F_0 over the last 75 ms. Flutter was at 0% from onset till 250 ms, then rapidly increased to 10% at 300 ms. It increased more slowly to a peak of 20% at 500 ms, after which it decreased to 0% at the end of the utterance. To model the irregularity of the glottal pulse often seen at the beginning and end of phonation, diplophonia started at 40%, and decreased to 0% at 75 ms; at 625 ms it began increasing again, returning to 40% at the end of the utterance. Amplitude of vocalization was 60 dB, aspiration noise was added into the glottal source at a constant 57 dB. The source switch (SS) was set to 4, leading to output of an unfiltered glottal source waveform.

Table 3: Vocal Tract Area Functions for Russian Vowels (from Fant 1960) Cross sectional areas (in square cm) for each section of the vocal tube, starting at the glottis (section length = 0.5 cm)

	Vowel		
Section #	[a]	[i]	[u]
1	2.6	3.2	2.6
2	1.6	2.6	2.6
3	1.3	2.0	2.0
4	1.0	2.0	2.0
5	4.0	8.0	10.5
6	2.6	8.0	10.5
7	1.6	10.5	10.5
8	1.0	10.5	8.0
9	0.7	10.5	8.0
10	0.7	10.5	5.0
11	0.7	10.5	3.2
12	1.0	10.5	1.6
13	1.3	10.5	1.3
14	1.6	10.5	1.0
15	2.0	10.5	1.0
16	2.6	8.0	1.0
17	2.6	8.0	1.6
18	1.6	6.5	2.0
19	3.2	4.0	1.3
20	4.0	2.6	1.6
21	5.0	1.3	2.0
22	6.5	0.7	2.0
23	8.0	0.7	2.0
24	8.0	0.7	2.6
25	8.0	0.7	3.2
26	8.0	0.7	5.0
27	8.0	0.7	6.5
28	8.0	0.7	8.0
29	8.0	1.0	10.5
30	8.0	1.3	13.0
31	6.5	1.6	13.0
32	5.0	3.2	13.0
33	5.0	4.0	13.0
34	5.0	4.0	10.5
35	5.0		5.0
36			2.0
37			0.3
38			0.3
39			0.7
40			0.7

To compute the formant frequencies for the four vowels and two vocal tract lengths used in this experiment, a more complicated approach was necessary than the simple tube model used in Experiment 1. Although a closed form solution for the formant frequencies is available for a two-tube model adequate for [i] and [a] (Fant 1960), [u] requires at least three tubes to be adequately modeled, prohibiting a simple solution. It was thus necessary to use numerical techniques to solve for the formant frequencies for these stimuli. The program TBFDA (for TuBe Frequency Domain Analyzer), written at MIT in FORTRAN (Henke 1966) and modified at Brown University by John Mertus, was used to analyze the cross-sectional area values given in Fant (1960) for each of the three (Russian) vowels [a], [i] and [u]. This program uses standard numerical techniques to solve a model of the vocal tract characterized by a series of sections of lossless tube, each with an independently specified cross-sectional area.

The cross-sectional areas for each vowel (from Fant 1960, p 115) are given in Table 3. In Fant's original data, the total vocal tract length (measured from glottis to lips) varied between vowels due to the effects of lip rounding for [o] and [u] and lip retraction for [i], as well as the longer flow path of [u]. For this experiment, however, the vocal tract shapes were scaled so that each vowel came from a vocal tract of exactly 15 or 18 cm in length. Thus, any effect of vowel on body size rating found in this experiment would be intrinsic to the vowel's spectral shape, and would NOT result from a difference in overall vocal tract length due to retraction or protrusion of the lips.

The formant frequencies and bandwidths calculated by the TBFDA program, using the input data given in Table 3, are shown in Table 4. The values for the schwa vowel [{] represent the output of the TBFDA program for a degenerate 1-tube vocal tract (included for comparison with Experiment 1). If the frequency of the fifth formant exceeded 5 kHz (the upper frequency practically allowed with a sampling rate of 11.025 kHz) for either vocal tract length, only the first four formants were used for both lengths. Bandwidths for the two different vocal tract lengths were kept fixed at the average value of the two vocal tract lengths (calculated separately for each vowel).

Table 4: Formant Center Frequencies (CF) and Bandwidths (BW) used in Expt 2

		Vowel							
		[a]		[i]		[u]		[{]	
VT length	Formant	CF	BW	CF	BW	CF	BW	CF	BW
18 cm:	1	596	70	214	70	245	61	474	69
	2	1010	82	2111	81	656	70	1421	76
	3	2335	116	2670	328	2623	67	2369	90
	4	3361	226	3435	139	4079	73	3318	111
	5	3975	108	4064	422				
15 cm:	1	707	70	254	70	291	61	565	69
	2	1196	82	2526	81	785	70	1696	76
	3	2787	116	3127	328	3148	67	2829	90
	4	3999	226	4105	139	4898	73	3963	111
	5	4763	108	4814	422				

In summary, the stimuli consisted of sixteen 700 ms vowel sounds which had a high or low fundamental frequency (100 or 150 Hz) and formants corresponding to a short or long vocal tract (15 or 18 cm) and one of four vowels ([a], [i], [u] or [{]), with no variation in vocal tract length between vowels.

Subjects

Subjects were 12 graduate students at Brown University with no medically-diagnosed hearing problems. All subjects spoke fluent English; all but two were native speakers of North American English (one was a native speaker of Hindi, the other Greek). There was no overlap between these subjects and those used in Expt. 1. Subjects were not paid for their participation.

Procedure

The stimuli were synthesized on an Apple Macintosh Quadra 800 computer using the Sensyn implementation of KLATTSYN88 and custom software written in C. The synthesized stimuli were transferred from the computer to a Panasonic SV-250 digital audiotape recorder via an Audiomedia II D/A board. Each of the sixteen sounds was presented six times resulting in 96 trials. One of two randomly-chosen orders was used for each subject. Playback was accomplished to each subject separately using the SV-250 DAT recorder and Realistic Pro-80 headphones (same signal in each ear), with volume adjusted to the subject's comfort. Two practice trials ([a] 15 cm, 150 Hz and [u] 18 cm 100 Hz) were run to allow the subject to ask questions before the experiment began. The instructions and rating task were the same in all other respects as that used in Expt. 1 (again, the endpoints of the rating scale were not fixed).

Experiment 2 Results

As in the first experiment, the subjects' ratings were approximately normally distributed, though subjects avoided the smallest rating almost completely this time. In general, although the same fundamental and vocal tract lengths were used, subjects rated the synthesized speakers from Expt. 2 as larger than those in Expt. 1: the mean rating was 4.6 (compared to 3.8 in the first experiment). Again, the number of scale points used varied from subject to subject (mean 5.5, range 3 to 7). The frequency distribution is shown in Fig. 9; the use of parametric statistics again seems warranted by the approximate normality of the data.

Figure 9: Frequency Distribution of Responses Combining all Subjects

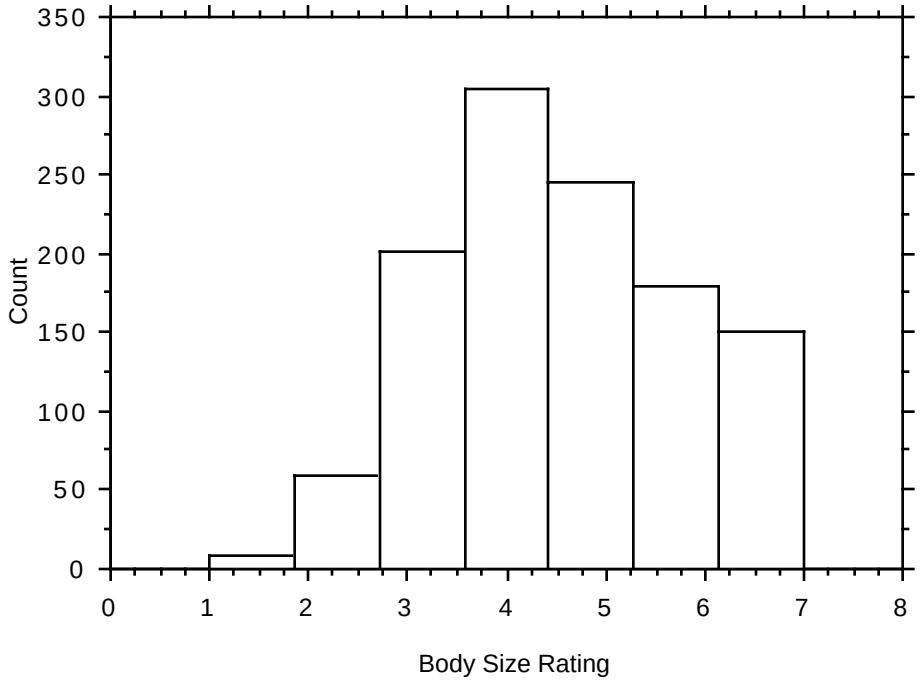


Table 5 shows the main effects and significant interactions for the 4-way repeated measures ANOVA used to analyze these data (illustrated in Figs. 10-12). The dependent measure was again body size rating (untransformed). As in Expt. 1, there was a clearly significant effect of both vocal tract length and fundamental frequency on body size judgments, but no effect of repetition. Thus Expt. 2 replicates the results of Expt. 1 with a different group of subjects, and somewhat different experimental protocol, and extends the results to three additional vowels which virtually cover the vowel space and have unevenly-spaced formants.

Table 5: 4-way Repeated Measures ANOVA for Expt. 2 (Significant Interactions Only)

Source	df	Sum Sqs	Mean Sq	F-Value	P-Value
Subject	11	187.42	17.04		
Vowel	3	34.65	11.55	3.48	.0721 NS
VT size	1	39.75	39.75	31.62	.0002 ***
Fundamental	1	1085.00	1085.00	47.31	.0001 ***
Repetition	5	1.28	.26	.26	.8228 NS
Vowel * VT size	3	15.49	5.16	4.86	.0236 *
VT size * Repetition	5	9.97	1.99	3.19	.0312 *

(N=12 subjects)

The interaction between stimulus repetition and vocal tract size, although significant, showed no interpretable pattern and will not be further discussed.

The interaction between vowel and vocal tract size resulted from the fact that vocal tract size had a much greater effect on body size judgments with the [i] vowel than with any of the other vowels (Fig. 12). This is interesting because it is consistent with the concept that the vowel [i] has a special status in vocal tract normalization (Nearey 1978, Lieberman 1984). Because [i] represents the extreme of oral constriction and pharyngeal expansion, it may represent an optimal vowel for vocal tract normalization since it has the lowest possible F1 and the highest possible F2. The effect found here suggests that subjects are more ready to interpret the variations in formant frequencies for [i] as changes in body size than they are with other vowels.

Figure 10: Change in Size Rating with Fundamental (Bars show Std. Error)

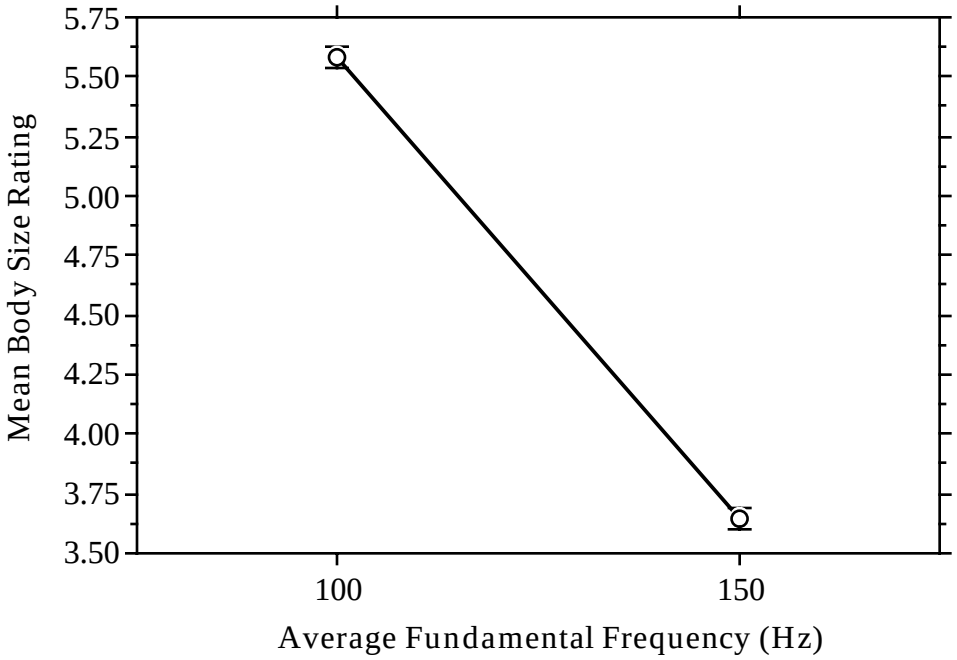


Figure 11: Change in Body Size Rating with Formants (Bars show Std. Error)

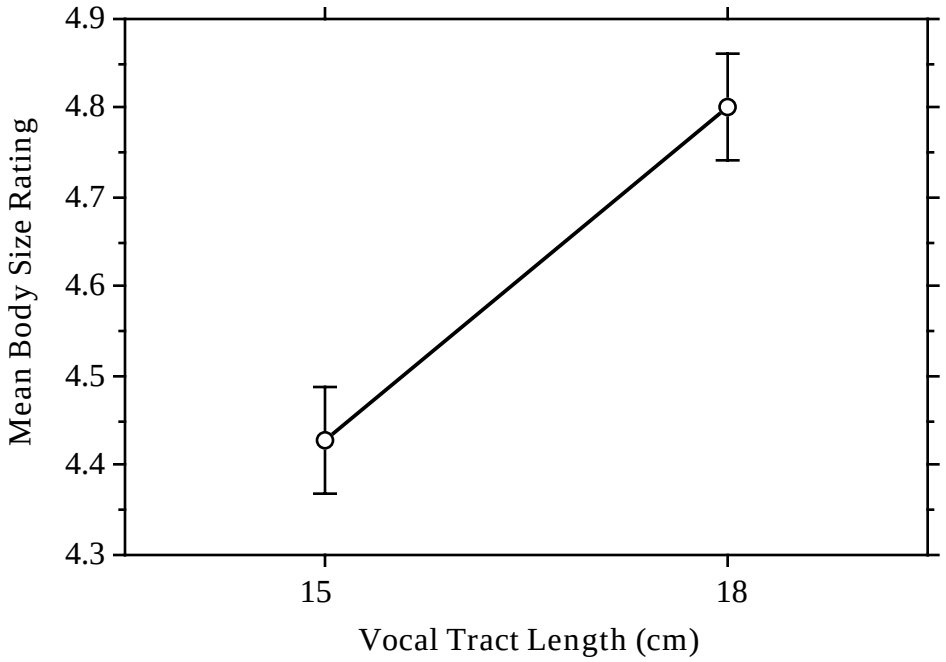
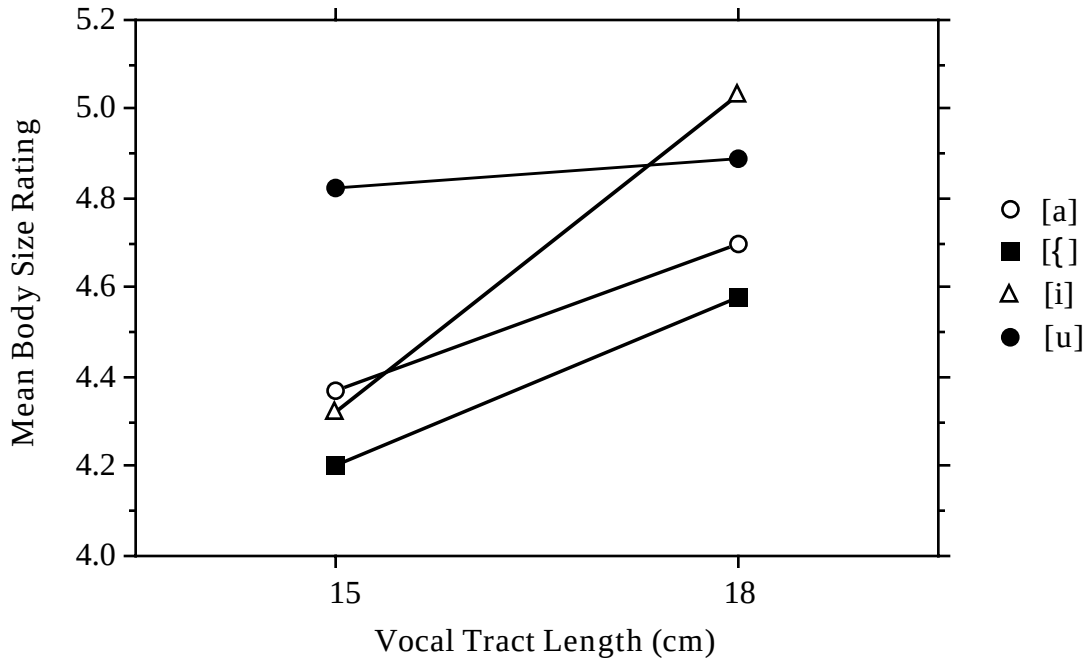


Figure 12: Interaction between Vowel and Formants



Experiment 3 Methods

Stimuli:

The stimuli used in this experiment were identical to those used in Experiment 2 in all but one respect. In the previous experiment, each of the vowels was generated with the synthesizer set to exactly one of two vocal tract lengths: 15 or 18 cm. In actual speech, a given speaker protrudes or retracts the lips while producing different vowels. For example, [i] is produced with the lips retracted, while [u] is produced with the lips rounded and protruded. This has the effect of lengthening (in the case of rounding) or shortening (in the case of retraction) the vocal tract, which in turn affects the formant frequencies.

In this experiment I allowed the vocal tract length to vary with vowel as it does in naturally-produced speech. The vowel [a] was taken to represent the unmodified vocal tract length (again, either 15 or 18 cm); this length was slightly shortened for [i] (to 14.59 or 17.16 cm, respectively), and lengthened for [u] (to 17.16 or 20.56 cm, respectively). These changes in vocal tract length simply reflect the unadjusted data from Fant (1960) (see Table 3).

The TBFDA program was again used to calculate the formant frequencies, using the input data given in Table 3. Two modifications of the TBFDA formant values were made to keep the stimuli between the two experiments as identical as possible. Although the longer vocal tract of [u] for this set of stimuli resulted in five formants for all vowels, unlike in Experiment 2, the same number of formant frequencies (i.e. five for [a] and [i], four for [u] and [{]) were used as in Experiment 2. Second, bandwidths from Experiment 2 were used for these stimuli (though the differences were only of a few Hz). The resulting formant frequencies and bandwidths are shown in Table 6.

All other aspects of the stimuli were identical to those in Experiment 2. The same glottal waveform files were used in the two experiments, so all aspects of pitch and timbre other than those deriving directly from the difference in vocal tract length were identical. The same synthesis and normalization procedure was used, using identical software. Playback was accomplished with the same DAT recorder and headphones. The same two random orders were used to present the stimuli. Identical instructions (on tape) and answer sheets were used.

Subjects

Subjects were 10 of the original 12 graduate students used in Expt. 2, tested four months after Expt. 2 to reduce the possibility of a learning effect (two of the subjects had moved out of the state in the interim). The reduction in sample size that resulted from this subject attrition worked against the vocal tract hypothesis, since it reduced the chance of results attaining significance in this experiment. (Two additional, randomly-chosen graduate student subjects who had not participated in any previous experiments were also tested, allowing a comparison with equal sample sizes). The stimulus presentation orders were the same as those used for the subjects in Expt. 2. As before, subjects were not paid for their participation.

Table 6: Formant Center Frequencies (CF) and Bandwidths (BW) used in Experiment 3

		Vowel							
Nominal		[a]		[i]		[u]		[{]	
VT length	Formant	CF	BW	CF	BW	CF	BW	CF	BW
18 cm:	1	596	70	220	70	216	61	449	69
	2	1010	82	2173	81	578	70	1351	76
	3	2335	116	2740	328	2297	67	2265	90
	4	3361	226	3536	139	3570	73	3192	111
	5	3975	108	4177	422				
15 cm:	1	707	70	262	70	256	61	531	69
	2	1196	82	2600	81	688	70	1599	76
	3	2787	116	3203	328	2752	67	2689	90
	4	3999	226	4221	139	4279	73	3799	111
	5	4763	108	4942	422				

Experiment 3 Results

Table 7 shows the main effects and significant interactions for the 4-way repeated-measures ANOVA used to analyze these data. Again, as in Expts. 1 and 2, there was a clearly significant effect of both vocal tract length and fundamental frequency on body size judgments, but no effect of the repetition of stimulus presentation. Unlike Expt. 2, where there was no significant main effect for vowel, there was a strongly significant effect for vowel in this experiment (shown in Fig. 13), as predicted by the vocal tract length hypothesis. The data closely follow the actual vocal tract lengths used in synthesis. Only the smaller ratings for the schwa vowel [{] are unaccounted for, since the same vocal tract length was used for it and [a]. However, since [{] was created with a degenerate, 1-tube vocal tract, and included primarily for comparison with the study in Chapter 1, the [{] ratings are of dubious importance.

Thus, the same subjects, exposed to an identical experimental protocol, show a strong effect for vowel if vocal tract length is allowed to vary with vowel, and the resulting body size ratings closely follow the prediction of the hypothesis that the difference in size judgments between vowels results from vocal tract length differences. (An equivalent analysis was run with 12 subjects, replacing the two subjects lost by attrition from Expt. 3, and the results were the same).

The interaction between vowel and vocal tract length discovered in Experiment 2 appeared again in this experiment, replicating the finding that the [i] vowel has a disproportionate influence on subjects' ratings of body size compared to the other vowels (Fig. 14). Figure 15 shows the three-way interaction between vocal tract length, F_0 and vowel which seems to result from the same phenomenon: the responses to the [i] vowel follow the same pattern as those for the other vowels, but are more extreme. The interaction between fundamental and vocal tract length (illustrated in Fig. 16) resulted from the effect of F_0 being more pronounced with the longer vocal tract. The latter two interactions were barely significant and will not be further discussed.

Table 7: 4-way Repeated Measures ANOVA for Expt. 3 (Significant Interactions Only)

Source	df	Sum Sqs	Mean Sq	F-Value	P-Value	
Subject	11	44.85	4.98			
Vowel	3	80.43	26.81	6.35	.0076	**
VT length	1	91.88	91.88	31.67	.0003	***
Fundamental	1	922.38	922.38	47.11	.0001	***
Repetition	5	2.61	0.52	0.69	.5333	NS
Vowel * VT length	3	16.23	5.41	7.86	.0040	**
VT length * Fund.	1	5.86	5.86	5.16	.0492	*
Vowel * VT * Fund.	3	8.84	2.95	3.33	.0475	*

(N=10 subjects)

Figure 13: Change in Body Size Rating with Vowel (Bars show Std. Error)

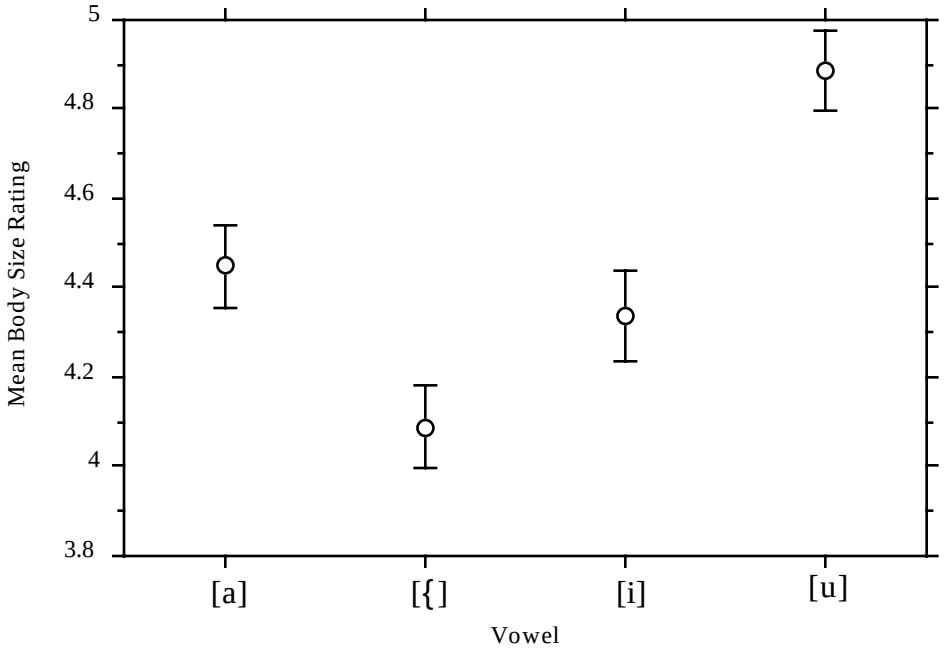


Figure 14: Interaction between Vowel and Vocal Tract Length

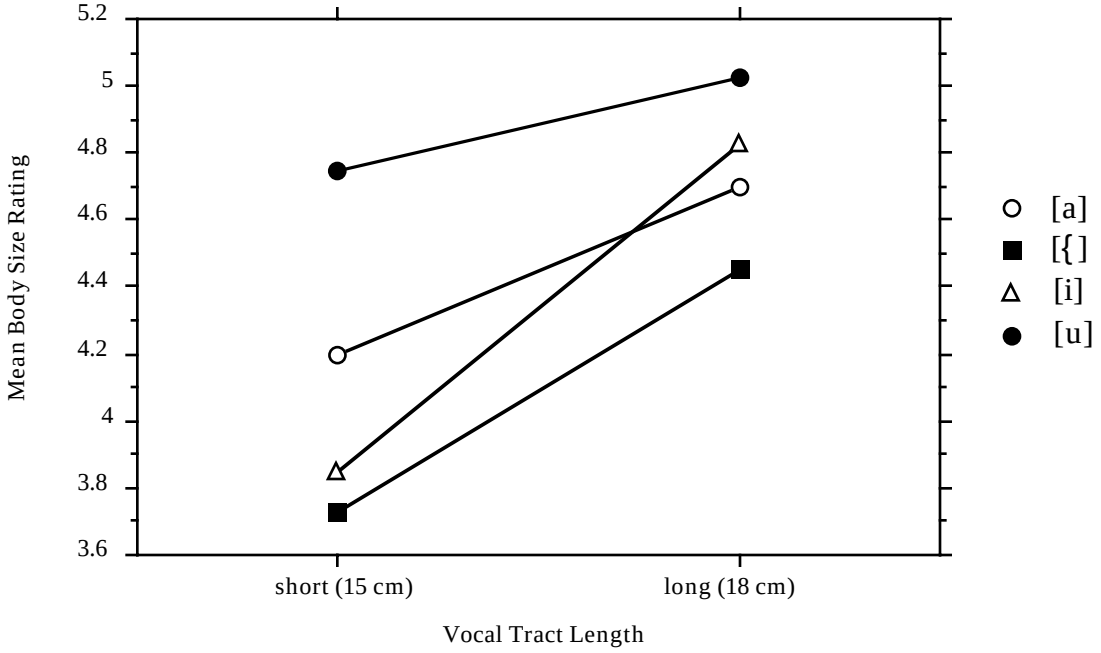


Figure 15: Interaction between F0, Vocal Tract Length and Vowel (Bars show Std. Error)

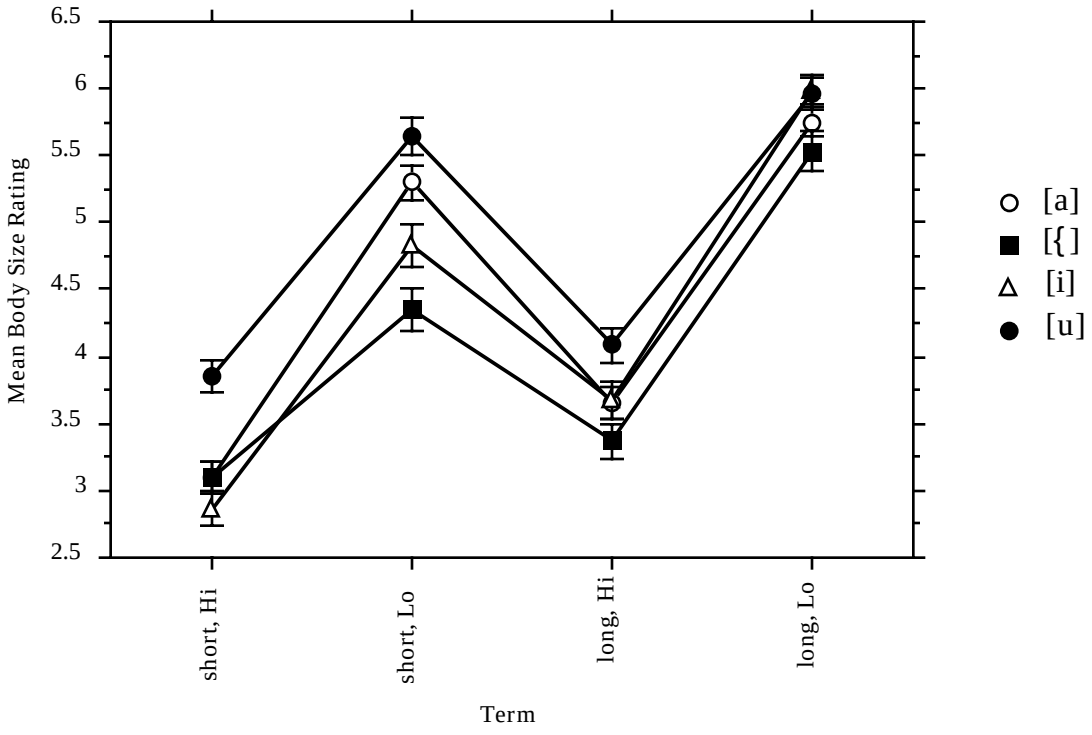
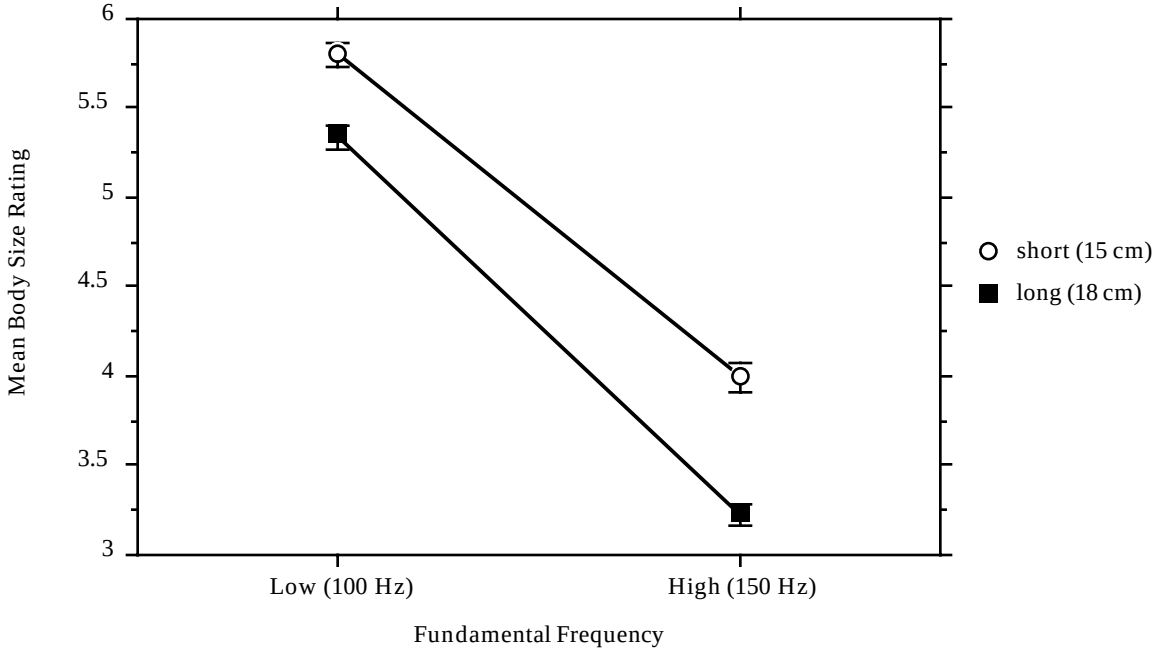


Figure 16: Interaction between F0 and Vocal Tract Length (Bars show Std. Error)



To test between the predictions of the "vocal tract length" hypothesis and the "F2 only" or "F2 minus F1" hypotheses, I performed a regression analysis of the values of F1, F2, and F2 - F1 (in Hz for each stimulus), and vocal tract length (in cm) against the body size ratings from both experiments (each presentation of the stimulus was averaged over subjects and used as a separate data point). Results are shown in Table 8. Only vocal tract length is a strong predictor, but both F1 and F2 show a slight but statistically significant relationship to body size rating. These variables, of course, covary with vocal tract length, and the fact that they are unable to drive a vowel-specific effect in Experiment 2 suggests that they play, at most, a secondary role in vowel symbolism when considered alone. Partial correlation coefficients for each of the three competing hypotheses (with the effect of vocal tract length removed) were less than 0.085 in all cases, while the correlation for vocal tract length in the same analyses exceeded 0.2. It is apparent from these data that the attempt made by "frequency coding" hypotheses to characterize a complex vowel spectrum by a single frequency value is unlikely to be successful.

Table 8: Regressions of single formant values and vocal tract length against mean body size ratings. Dependent variable: Body size rating.

Independent Variable	r Value	F Value	p Value
VT Length (cm)	.260	13.76	.0003
F1 (Hz)	.154	4.631	.0327
F2 (Hz)	.161	5.059	.0256
F2 minus F1 (Hz)	.107	2.22	.1380

Experiments 2 and 3: Discussion

The results of these two experiments provide strong support for the hypothesis that vowel symbolism for size is accounted for by differences in vocal tract length which accompany the production of the different vowels. The same group of subjects showed a strongly significant vowel/size effect when the vocal tract length producing the vowel varied (as in normal speech), but showed no effect when the vocal tract length was held constant for the different vowels. Experimental protocol, stimulus order, playback conditions and all aspects of the stimulus except vocal tract length were constant across the two experiments. These results suggest that other hypotheses which have been proposed to explain vowel/size symbolism are inadequate.

One explanation for vowel/size symbolism would be that it is an example of conventional phonetic symbolism (e.g., Taylor & Taylor 1962, Taylor 1963, Taylor & Taylor 1965). This hypothesis is refuted by the data presented here, because the phonetic identities of the vowels were the same in the two experiments. If the effect were based purely on a conventional association between vowel identity and size, it would appear equally strong in both experiments.

The other two hypotheses tested in these experiments were the "F2 alone" hypothesis and the "F2 minus F1" hypothesis (Woodworth 1991, Ohala 1984). Both of these hypotheses predict approximately equal results in Experiments 2 and 3, and are unable to explain the observed difference. When the predictions of each of the hypotheses are tested against the data from both experiments, the vocal tract length hypothesis provides a good fit for the data ($p < .0005$), while the "F2 minus F1" hypothesis provides negative results ($p > .10$). Both F1 and F2 show some correlation with body size ratings ($p = .033$ and .026, respectively), so these data provide no support for the contention that F2 plays a special role.

Thus the results of Experiments 2 and 3 provide strong support for the idea that vowel symbolism for size derives from the variations in vocal tract length associated with the production of different vowels. A remaining potential weakness in the argument results from the possibility that a learning effect occurred during Experiment 2 which, four months later, resulted in the effect observed in Experiment 3. This seems unlikely, not only because four months would be a long period of time for such an effect to persist, but also because there was no evidence of a learning effect during the course of the 96 trials in Experiment 2 (see Table 5). The inevitable attrition of subjects which resulted from the long period of time between test and retest worked against the vocal tract length hypothesis, since the resulting reduction in sample size made it less likely that a statistically-significant effect for vowel would be found.

However, in order to eliminate the possibility of learning being responsible for the effect seen in Experiment 3, and to strengthen and extend those findings by demonstrating the same effects with slightly different stimuli and presentation techniques, a fourth experiment was conducted with a new set of subjects.

Experiment 4 Methods:

Stimuli:

The stimuli used in this experiment were identical to those used in Experiment 3 in most respects. The same glottal waveform and synthesis techniques were used as with Experiments 2 and 3. The only difference is that I relaxed the stringent controls which were used to keep the stimuli for those experiments identical in terms of formant bandwidths and number of formants (which resulted in the formant values in Experiment 3 being slightly different than those calculated with the TBFDA program). Thus, the stimuli used in this experiment were exactly as predicted by the TBFDA calculations.

The resulting formant frequencies and bandwidths calculated by the TBFDA program, using the input data given in Table 3, are shown in Table 9. All vowels had five formants below 5 kHz. Again, bandwidths for the two different vocal tract lengths were kept fixed at the average value (calculated separately for each vowel). In the case of [u], there was a slight high-frequency ringing due to overly narrow bandwidths for the highest 3 formants. This was corrected for by substituting the average bandwidth from the other three vowels for the top three formant bandwidths in [u]. All other aspects of the stimuli were identical to those in Experiment 3.

Table 9: Formant Center Frequencies (CF) and Bandwidths (BW) for Expt. 4

		Vowel							
Nominal		[a]		[i]		[u]		[{]	
VT length	Formant	CF	BW	CF	BW	CF	BW	CF	BW
18 cm:	1	596	69	220	65	216	62	449	71
	2	1010	84	2173	100	578	65	1351	126
	3	2336	118	2740	345	2297	220	2265	198
	4	3362	225	3536	139	3570	210	3192	268
	5	3977	111	4177	408	3909	288	4131	347
15 cm:	1	706	69	262	65	256	62	531	71
	2	1195	84	2600	100	688	65	1599	126
	3	2784	118	3203	345	2752	220	2689	198
	4	3996	225	4221	139	4279	210	3799	268
	5	4758	111	4942	408	4682	288	4927	347

Subjects

Subjects were 12 undergraduate students at Brown University with no medically-diagnosed hearing problems. Subjects were all native English speakers. There was no overlap between these subjects and those used in the first two experiments. Subjects were paid for their participation.

Procedure

As in the previous experiments, each of the sixteen sounds was presented in a randomized order six times, resulting in 96 trials. Playback was accomplished to groups of 4-8 subjects using the SV-250 DAT recorder and Bose Roommate II loudspeakers (same signal from each speaker), in the same room and conditions used in Expt. 1. Two practice trials ([a] 15 cm, 150 Hz and [u] 18 cm 100 Hz) were run to allow the subject to ask questions before the experiment began. The instructions and rating task were the same in all other respects as that used in the first three experiments (again, the endpoints of the rating scale were not fixed).

Experiment 4 Results

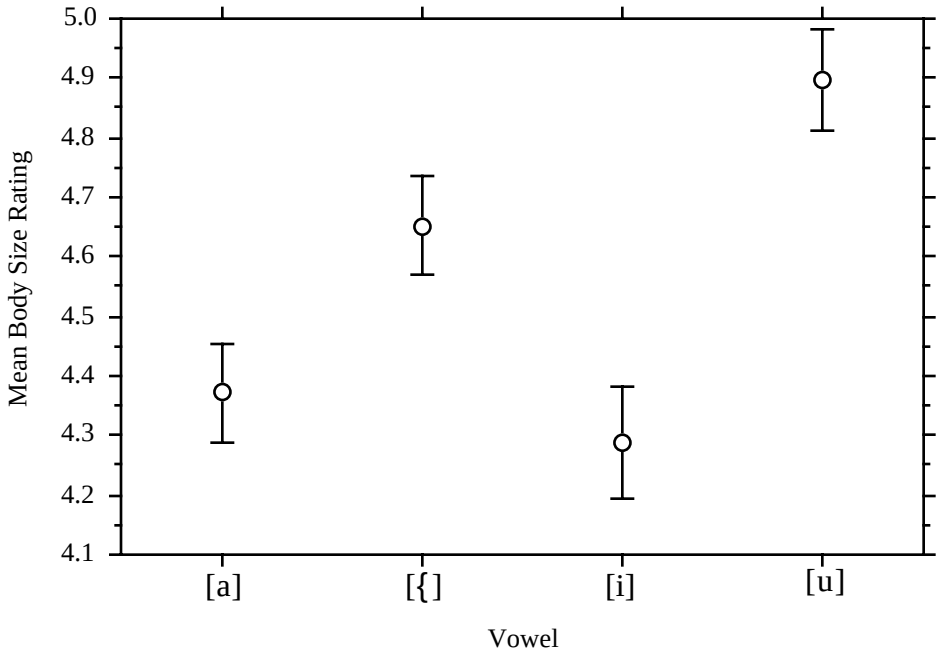
Table 10 shows the main effects and significant interactions for the 4-way repeated measures ANOVA used to analyze these data. As in the first two experiments, there was a clearly significant effect of both vocal tract length and fundamental frequency on body size judgments. As in Experiment 3, a highly significant effect for vowel was found, although in this case the subjects had never been exposed to the task before. Experiment 4 thus replicate the findings of the first two experiments with yet another group of subjects, and an additional permutation of experimental protocol.

Table 10: 4-way Repeated Measures ANOVA for all data (Significant Interactions Only)

Source	df	Sum Sqs	Mean Sq	F-Value	P-Value
Subject	11	401.146	36.468		
Vowel	3	66.396	22.132	8.889	.0008***
Vocal Tract Length	1	22.222	22.222	15.153	.0025**
Fundamental	1	734.722	734.722	38.385	.0001***
Repetition	5	27.323	5.465	4.370	.0292*
Vowel * VT Length	3	22.757	7.586	5.282	.0114*
VT Length * F_0	1	22.222	22.222	11.522	.0060**
(N=12 subjects)					

Figure 17 shows the mean body size ratings for each vowel. Once again, the body size ratings follow closely the predictions of the vocal tract length hypothesis, and only the degenerate one-tube schwa stimulus behaves strangely. Although subjects were not asked to rate the naturalness of the different vowels, informal listening suggests that the schwa was a much less natural-sounding stimulus than the other three vowels; the deviation in its size rating may be a reflection of this fact.

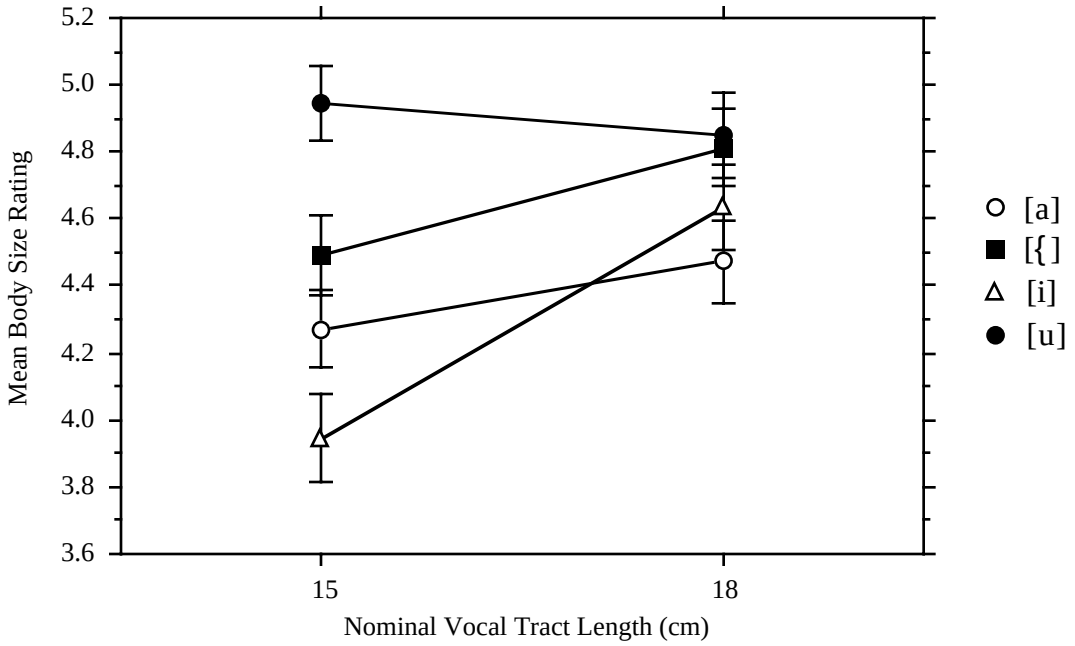
Figure 17: Change in Body Size Rating with Vowel (Bars show Std. Error)



As in the previous two experiments, there was a significant interaction between vowel and nominal vocal tract length (Fig. 18). The lack of increase in size judgments with vocal tract length for the [u] vowel is surprising, but the more outstanding feature is again the larger effect of vocal tract size with the vowel [i] compared to other vowels. These data are consistent with the proposal of Nearey (1978)

and Lieberman (1984) that [i] functions as a "supervowel" which plays a special role in vocal tract normalization and hence speech perception.

Figure 18: Interaction between Vowel and Formants (Bars show Std. Error)



General Discussion

How do human languages achieve their expressive adequacy? The range of topics which can be explored using language is truly astounding: we can discuss, with great precision but also considerable economy of expression, the operation of an automobile engine, the personalities and interactions characterizing a family, or the meaning of abstract concepts like Truth or Beauty. Linguistic interactions can range from almost purely abstract (e.g., describing the process of logging onto a computer), or their message can be mainly emotional or social (most of us have experienced arguments whose main purpose is venting of anger or establishment of dominance, but which are nominally "about" some mundane topic like taking out garbage or leaving the cap off toothpaste). In many ways, the most remarkable aspect of language is its ability to flexibly and transparently span such extremes. Although it is popular to class the latter topics as "paralinguistic" or "primitive," and hence of little importance, the vast majority of utterances have at least some emotive or social component forming part of their communicative impact. It seems likely that phonetic symbolism plays an important role in the expressive use of language; I suggest that the isolated and relatively tractable nature of the phenomenon makes it a promising foothold in an otherwise treacherously slippery area.

The remarkable fact is that language is flexible enough to convey both abstract and social/emotional messages, often simultaneously, without any apparent decrement in its communicative accuracy. It is of course true that artificial languages such as mathematics or formal logic are more precise and consistent than any natural language, and perhaps the case that pantomime or music are capable of more transparently expressing emotional or social messages (though I personally doubt this),

but none of these types of communication have been adopted for use in day-to-day discourse. Even skilled mathematicians or musicians still make extensive use of natural language in their interactions with one another.

The ability of language to communicate both social/emotional and abstract messages has some interesting implications for the evolution of language. Sometime in the not too distant past (it doesn't matter when, though current estimates are about a half million years ago (Lieberman 1991)), our hominid ancestors lacked language as we know it. But judging both from the archaeological evidence and contemporary non-human primates, they probably led relatively complex social lives and used some system of gestural and vocal communication to mediate their interactions with one another. The uses to which this communicative system were put may seem relatively mundane (and indeed little different from those of other social animals): choosing and bonding with mates, forming alliances, establishing cooperation amongst community members, establishing dominance, deterring cheating and exerting social control. But although modern human language has added to these "primitive" capabilities the ability to discuss abstract counterfactual constructs, transfer richly elaborate structures of knowledge from parent to child, and even discuss language itself, modern humans still use language to achieve all of the "primitive" goals cited above. Although accounting for these added capabilities is clearly interesting and challenging, an account of language which focuses on them alone and ignores the incredible wealth of social and emotional communication also mediated by language will simply be descriptively inadequate.

There are several reasons to believe that phonetic symbolism can provide significant insights into the expressive adequacy of language. A major task in the study of language is a description of the interface between abstract linguistic structure and meaning (in its fullest sense, including both denotative and the full range of emotional and social connotative meaning). Much of the progress made in understanding linguistic structure in this century was made possible by the simplifying assumption of the arbitrariness of the sign, and more particularly the meaninglessness of the basic units of phonology: phonemes and features. It is a pleasant surprise to find that these structural units can actually play a significant role in the expressive capacity of language, providing a well-defined phenomenon with which to explore the articulation between structure and meaning. This allows us to clarify (and even quantify) the concept of "expressive adequacy" in a phenomenon which is both broadly relevant (being found in a broad sample of the world's languages) but also specific and circumscribed enough to allow thorough, careful, experimental investigation.

In the sections that follow, I will outline some aspects of language study where phonetic symbolism could provide interesting insights, and suggest some possible approaches to its quantification and experimental investigation. I will suggest two areas where phonetic symbolism seems especially likely to be important: in language change and in child language acquisition. Of course, these comments should be taken as suggestive, and a review of all the possible implications of phonetic symbolism in the world's languages would take us far afield (Malkiel's 1990 book gives a sense of the broad implications of "phonosymbolism" for historical linguistics; Fordyce's 1988 thesis reviews the types and extent of phonetic symbolism in the world's languages in some detail).

Phonetic Symbolism and Historical Linguistics

Where do words come from? Why do some words persist and flourish while others disappear? What kinds of forces cause the shifts of meaning and sound which continually flow through natural language? A complete theory of historical linguistics would come to grips with questions like these. Several historical linguists (Jespersen 1922, Malkiel 1990) have proposed that phonetic symbolism may play an important role in language change. For example, Jespersen suggested that languages become more expressively adequate during their development because languages with consonance between sound and meaning outsurvive those which lack such consistency. The Latin "parvus" (small) was an ill fit to its meaning in terms of vowel symbolism for size; it has evolved in modern Romance languages to French "petite" and Spanish "pequeño", which are both more sound symbolic. In English,

the word “small” (which does not display vowel/size symbolism) is relatively technical and inexpressive, while “little” (which does show vowel/size symbolism) is more expressive, but it is outdone by the further extensions “tiny”, “teeny” and “teeny weeny”, which grow progressively more sound symbolic. (Interestingly, [tini] is the ME pronunciation of “tiny”, which was changed to [tayni] during the Great Vowel Shift; did this necessitate the coinage of “teeny”?)

Such ideas are not new. But the vocal tract length explanation of vowel/size symbolism allows us to evaluate them much more rigorously than was possible before. I predict that words for “small” and “large” in the world’s languages will use vowels produced with short and long vocal tracts (respectively) synchronically, and that words which don’t obey this relationship will change in that direction diachronically. Words which obey sound-symbolic rules will be resistant to change. (Of course, the principles of arbitrary and regular sound change play an unquestionably central role in language change, explaining the fact that all languages haven't grown more and more sound-symbolic). These predictions can easily be tested in a rigorous fashion (see, for example, Appendix A, where I test for synchronic vowel/size symbolism in a broad sample of the world’s languages, with strongly positive results). We can thus explore what aspects of a word determine its “fitness” during the development of a language, and even gain insight into the murkier question of how new words are coined in the first place, and what factors influence their subsequent fates. It seems likely that the role of phonetic symbolism in these respects might be considerable; the account of vowel/size symbolism developed here provides a basis for empirical exploration of this role.

Child Language Acquisition

A major task facing the language-learning child is the acquisition of a huge vocabulary in a relatively short period of time: approximately 5000 words per year or thirteen words per day during the early school years (Miller and Gildea 1987). Furthermore, vocabulary size continues to increase through the college years, giving the average college student a 50,000 word receptive vocabulary (Just & Carpenter 1987). Until college, it is rare for children to learn words via explicit definitions, and instead meanings are typically inferred from context and usage (Miller & Gildea 1987). The enormous task of learning all these word meanings will be made less difficult if there is any degree of sound symbolism in the language, because of the consequent reduction of the uncertainty involved in learning new words. Either mimetic or conventional sound symbolism would provide a useful aid, though conventional symbolism will only help when learning a word in the same sound-symbolic cluster as a previously-acquired term.

Jespersen (1933) suggested that sound symbolism plays a considerably stronger role in children’s language than in adults’, and Sapir’s experimental investigations (1929) were consistent with this idea, showing that vowel symbolism for size was in place in English-speaking children at age 11 in much the same form as for adults. Sapir’s nonsense word paradigm, which has been widely and successfully exploited in the exploration of sound symbolism in adults (e.g., Johnson, Suzuki and Olds 1964, Taylor and Taylor 1962, Tarte & Barrit 1971, Tarte 1974) can be easily applied with children to investigate the role of phonetic symbolism in language acquisition. One technique would use nonsense words for different sizes of toys which either exhibit vowel/size symbolism (e.g., large toys have [u] or [o] in their name, while small toys contain [i] or [e]) or do not (having either random vowels or a non-symbolic pattern). If sound symbolism is playing a role in vocabulary learning, we would predict that the sound-symbolic words will be learned faster and remembered better. (In this experiment, the words should be presented by a researcher blind to the purpose of the study, or better yet presented with a speech synthesizer to avoid the use of non-phonetic cues).

Conclusion

In this chapter I reviewed the evidence that vowel symbolism for size exists in English, and that it is a type of mimetic symbolism and not purely conventional. I offered a new psychoacoustic explanation

of why this is the case, contrasting my hypothesis with several others from the literature. Then, using computer-synthesized vowels and a psychoacoustic task having nothing to do with word-meaning, I tested among these hypotheses. Only one hypothesis was able to account for the pattern of results: that vowel symbolism for size derives from the variation in vocal tract length accompanying the production of the different vowels. These experiments represent evidence for a physically-based explanation of one type of sound symbolism, but they also represent a general paradigm for the rigorous study of sound symbolism. In the final section, I explored some implications of phonetic symbolism for language change and language acquisition, and outlined some experimental approaches for extending the results reported here. I conclude that phonetic symbolism may play a much more important role in natural language than heretofore recognized, and that furthermore it is well-suited to rigorous empirical investigation. These experiments represent one small step in that direction.

Chapter 3: Vocal Tract Length in Non-Human Vocal Communication

In the previous two chapters, I have presented evidence that human beings use vocal tract length, via its acoustic correlate of formant dispersion, to estimate the body size of a speaker. This information is probably used in vocal tract normalization, a process fundamental to speech perception. It also appears to account for vowel/size symbolism, the best-documented type of sound symbolism, in which vowels produced with short vocal tracts are used in words for small things (and conversely, vowels produced with long vocal tracts are used to signify large things). Thus, the acoustic correlates of vocal tract length appear to play an important role in human language. In this chapter, I will address the question of whether or not these effects are unique to humans or instead result from more ancient perceptual mechanisms which we share with other animals.

There are several reasons to suspect that animals might use formant dispersion to determine vocal tract length and thus estimate body size. The first is that this acoustic cue has been available for eons: all vertebrates have some sort of vocal tube (minimally the oral cavity), which inevitably filters the laryngeal signal and creates formants in the output which provide a direct cue to the length of the vocal tract. Also, we know from studies of animal psychophysics that rhesus macaques are able to perceive formant frequencies, and indeed do so marginally better than human beings (Sommers et al., 1992). Given that body size plays an extremely important role in animal ecology and social behavior (Archer 1988, Parker 1974, Peters 1983, Schmidt-Nielsen 1984), there would be strong selective pressures to make use of these formant cues to assess the body size of potential mates or competitors.

Another clue that vocal tract length may have played a role in non-human vocal communication comes from a rather puzzling primate facial display variously known as the "grin face" or "bared-teeth display" (Andrew 1963, van Hooff 1972). This display, probably analogous to the human smile or laugh, involves a retraction of the mouth corners, revealing the teeth. The display is typically used in contexts of affiliation or appeasement, frequently being performed by subordinate animals to avoid or ameliorate aggression by a dominant individual. Treated as a visual sign, the use of a display which reveals the teeth (including in most primates a fearsome set of canines) to avoid aggression seems surprising, if not inexplicable. Indeed, a number of aggressive displays such as the macaque threat yawn (van Hooff 1967), hippo threat display (Walther 1977) and solenodon open mouth threat (Poduschka 1977) reveal the teeth in what is clearly a show of weaponry. Why then is the teeth-displaying "fear grin" nearly universal as a submissive display among primates? One plausible resolution to this puzzle involves the fact that the fear grin is often accompanied by vocalizations. The acoustic effect of lip corner retraction is to shorten the effective length of the vocal tube (Ohala 1984). If animals use formant dispersion and vocal tract length to gauge body size, this would have the perceptual effect of making the animal seem smaller, and thus less threatening. This is of course precisely the message that we would expect a submissive organism to project. This hypothesis, termed "the acoustic origin of the smile" by Ohala (1980), obviously requires that listeners perceive formant dispersion and use it as a cue to body size.

Finally, animal sensitivity to formant dispersion might be expected based simply on the Darwinian point of view positing the continuity and gradualness of evolutionary change (see Chapter 4). Humans share much of our anatomy, physiology and behavior with our animal cousins, and everywhere we look (whether at a genetic, anatomical or ethological level of analysis) we find homologies between humans and other species. This is no less true of the biology of speech than in other realms (e.g., the anatomy of our larynges, the neural and acoustic basis of phonation, and the basic anatomy and physiology of the auditory system are clearly very similar in humans and many other species). From this perspective it would be surprising if there were no homologies in slightly more abstract perceptual functions underlying our communication system.

Given this *a priori* support, it seemed worthwhile to seek experimental evidence that animal communication systems make use of vocal tract length, and its acoustic correlate of formant dispersion, as a cue to body size. Such evidence could take several forms. The simplest might be a simple

examination of vocal tract length and formant dispersion in organisms of different body sizes. If we discovered a significant correlation in such a study, it would be suggestive: it would show the cue is available, so that listeners could theoretically use this cue to judge body size. But it would obviously not constitute evidence that they do.

Taking a different tack, we could use operant conditioning techniques to repeat Experiment 1 of this thesis, this time with animal subjects: asking them to tell us which of two stimuli sounds larger. This would first involve initially training animals to make size judgments in the visual domain (e.g., train them to push a button indicating which of two objects is larger). Once the subjects reached some criterion of adequate performance, we could then substitute auditory stimuli. If the animals were able to transfer the task from the visual to the auditory domain, we could use computer-synthesized sounds to address the formant dispersion cue. However, this experiment would suffer from a similar problem as the previous study: it would show that the animals possess the perceptual apparatus necessary to judge body size from formant dispersion, but not that they actually make use of it in their communication systems.

To address this problem, we can take advantage of the fact that many species share the human ability to modify vocal tract length via lip maneuvers. Most mammals can shorten the effective acoustic length of the vocal tract by retracting the lip corners (in animals with pronounced snouts, such as dogs, this probably reduces the effective vocal tract length by half or more). Many mammals (particularly primates) can also lengthen the vocal tract by extending and rounding the lips (lip protrusion and lip rounding are both effects of contracting the same muscle, the orbicularis oris, and have the same acoustic effect of lengthening the vocal tract: Stevens and House 1955). If listeners use the formant dispersion as a cue to body size, we expect vocalizers to make use of the range of variation in vocal tract length that they have at their disposal differently in different behavioral contexts. For example, in situations where we might expect an individual to benefit from a projection of small body size (e.g., to seem small and unthreatening), we would predict lip retraction. This is precisely the prediction which led to Ohala's (1980) proposal of the acoustic origin of the smile. Of course, it is possible that listeners make use of formant dispersion, but vocalizers do not manipulate the cue. However, it is widely agreed that the details of animal displays have evolved in accordance with their effects on receivers (Andersson 1980, Krebs and Dawkins 1984), so a demonstration that vocalizers manipulate the cue in the expected way would constitute strong evidence that listeners are sensitive to it.

Unfortunately, there is a missing link in the approach outlined above: how do we know which behavioral contexts favor "seeming" small? Although the example of submission may be relatively clear, there are many situations in which this is not true For example, when a frightened infant runs to its mother, what behavior would we expect from the mother? It seems equally plausible that she would do best to seem small (and unthreatening) or large (and thus powerful, safe and comforting). I can see no convincing reason to expect one or the other. Clearly what is needed is some independent means of assessing what body size a vocalizer is attempting to project.

In this study, I used the degree of fur erection or piloerection ("raising the hackles") as a measure of projected body size. Piloerection, though originally involved in temperature regulation (Morris 1956), is widely agreed to have further evolved as a means of increasing the apparent body size of the displaying animal (e.g., Goodall 1986, Marler 1968, Morris 1956, Wilson 1972). This display has no acoustic consequences, and thus provides a suitable independent measure of projected body size for this study.

I observed the social behavior of captive white-faced saki monkeys (*Pithecia pithecia*), and estimated the degree of lip protrusion during vocalization, and searched for a correlation between lip protrusion and piloerection. Such a correlation would provide support for the hypothesis that these monkeys elongate their vocal tracts, thus increasing their acoustically-projected size in those situations in which they increase their visually-projected size by piloerection. Sakis are ideally suited for this

experiment: they vocalize frequently and have an extremely thick coat of fur which makes piloerection relatively obvious. Saki vocal behavior is undescribed and very little is known about their behavior in general (though see Olivera et al. 1985, Dugmore 1986). *P. pithecia* is a neotropical species which lives in dense rainforest environments in South America, suggesting that acoustically-interacting conspecifics will frequently be in situations of limited visual contact. This would presumably encourage the development of mechanisms for assessing, and manipulating, acoustically-projected body size.

Methods

Animals and Housing

8 saki monkeys (*P. pithecia*) at the Roger Williams Park Zoo in Providence served as subjects, consisting of three adult pairs and two juveniles in one large room. No data were obtained from either juvenile, or one of the adult females, so the final sample consisted of five adult monkeys (3 males) between the ages of five and six years. All of the monkeys in the colony were born in captivity and were mother-reared. Each adult pair was housed in a separate 2.1 x 3.1 x 3 m indoor cage; two of these cages also housed one juvenile child of the pair each (one of each sex, each one year of age). Cages were separated by an opaque plastic barrier, preventing visual contact between cages. Monkeys had free access to water and were fed a mixture of Purina New World Monkey Chow and Canned Primate Diet daily before observations began, and mixed fruit, greens, vegetables, nuts, seeds, hard-boiled eggs and mealworms after observations. Temperature in the colony room was maintained between 26°-28° C throughout the course of the study.

Video Clips

Animals were filmed during their normal daily behavior and during several mild manipulations to obtain a wide variety of behaviors and vocalizations. The normal daily behaviors filmed included feeding, self-grooming, nursing, other affiliative behaviors, and spontaneous aggression directed at either the observers or monkeys in neighboring cages (no aggression was observed between inhabitants of the same cage). The manipulations, designed to increase aggressive behavior, included a 0.5 m movement of one cage to give neighboring monkeys a limited view of one another, and holding up a small 0.75 x 0.5 m mirror to the cage wall Both of these manipulations resulted in a substantial increase in aggressive behavior in pilot studies.

Videotaping took place in the morning when the monkeys were most active, before and after their normal feeding time (all filming was performed by the author between 8 AM and 12 AM, on five days between Aug. 26 and Sept. 3, 1993). The video camera was a hand-held Panasonic Pro-Line AG-195P (using Fuji A/V Pro, VHS format videotape) using ambient light (provided by two large windows and overhead fluorescent bulbs). Vocalizing individuals were followed with the camera to allow close-up filming of their mouths, resulting in one hour and 36 minutes of usable footage. All video footage was then reviewed; if it was possible to observe an animal's mouth as it vocalized, the entire sequence of that animal's behavior leading into and following the vocalization was given a track number and copied onto a Fuji H471S Super-VHS master tape. A total of 38 short clips resulted, ranging in length from 6 to 50 seconds (mean clip length: 21 secs).

Ratings

These video clips were then rated for piloerection and lip protrusion by raters naive to the hypothesis under test. In order to perform ratings, two experimental tapes (Fuji A/V Pro, VHS format) were constructed from the master tape in two different random orders. Before each clip, a black screen faded into a label reading "TRACK N", where N was the clip number. The total duration of black screen + label, which constituted the entire pause between clips, was 8 seconds. Each tape had two example clips at the beginning, giving examples of the extremes of piloerection or lip protrusion, as appropriate (these clips were not reused during the actual trials). Raters had a remote control to allow

them to pause or review the clips at their leisure. For both films, raters were asked to always rate the maximum amount of the behavior seen at any point in the clip.

The first tape was used to score piloerection. Raters were asked to assign a number to each clip between one (corresponding to no piloerection) and five (full piloerection), along with a confidence rating from one to five (5 = certain, 1 = guessing). At the beginning of this tape, two short example clips were provided as examples of "no piloerection" and "full piloerection". Because I suspected that changes in posture such as shoulder hunching or adopting a convex arch in the spine might affect apparent body size independent of piloerection (e.g., Kenyon 1969, Moynihan 1967), raters were also asked to rate "if they thought the animal was trying to appear larger, for example by hunching the shoulders or arching the spine", again using a five point scale (5 = they were certain the animal was trying to look larger). These two rating tasks were performed concurrently on the first tape. The sound from the video was turned off to prevent acoustic cues from having any effect on these ratings.

The second tape used the same 38 clips as the first, arranged in a different random order. For this tape raters were asked to estimate the maximum degree of "lip protrusion and rounding" exhibited by the animal while vocalizing, using a five point scale where five meant full protrusion and one meant slight retraction of the corners of the mouth. Due to the anatomy of the orbicularis oris muscle, rounding and protrusion of the lips occur in tandem, and both have the same acoustic effect on vocal tract length (Stevens & House 1955). Typically, rounding was most visible from a frontal view of the monkey's face, while protrusion was most evident from a side view. Drawings of the monkeys faces were provided showing fully rounded, relaxed and retracted mouths, in front and in side view. Again, two short example clips were played at the beginning of the tape to orient the raters to the endpoints of the scale. During this rating task, sound was turned on to allow the raters to determine whether or not the animal vocalized.

Raters were given worksheets on which to make their answers. One set of raters received the piloerection rating task first, the other the lip protrusion task (order randomly determined). The raters were four Brown University graduate students. The raters reported that they had no prior experience with saki monkeys, and were ignorant of the purpose of the study and the hypothesis under test.

Results

Ratings of piloerection and lip protrusion were tabulated for each video clip. The mean ratings (combining the four individual raters) for each clip for the piloerection and lip protrusion/rounding variables were then compared using a Spearman Rank Correlation analysis. The two variables were significantly correlated (N = 38 film clips, Spearman rho (corrected for ties) = .579, p = .0004). The results are plotted in Fig. 19.

Spearman Rank Correlation analysis of the two ratings for each rater individually are given in Table 11. Piloerection and lip protrusion were significantly positively correlated for three of the four raters individually.

Inter-rater reliability was relatively low for the piloerection ratings (correlations ranged from .228 - .755), and higher for the lip protrusion/rounding ratings (.534 - .726). All of these inter-rater correlations were significant except one (p = .0001 to .0321 for the 11 significant correlations, p = .1694 for Rater 1 vs. Rater 2 piloerection). These correlations reflect the rater's reported difficulty gauging the degree of piloerection in many of the clips, typically due to movement by the animals. All raters found lip protrusion/rounding to be relatively easy to judge; this is not surprising since the initial selection of the film clips was based on the visibility of the mouth.

Figure 19: Mean Piloerection Ratings vs. Mean Lip Protrusion Ratings (combining all raters)

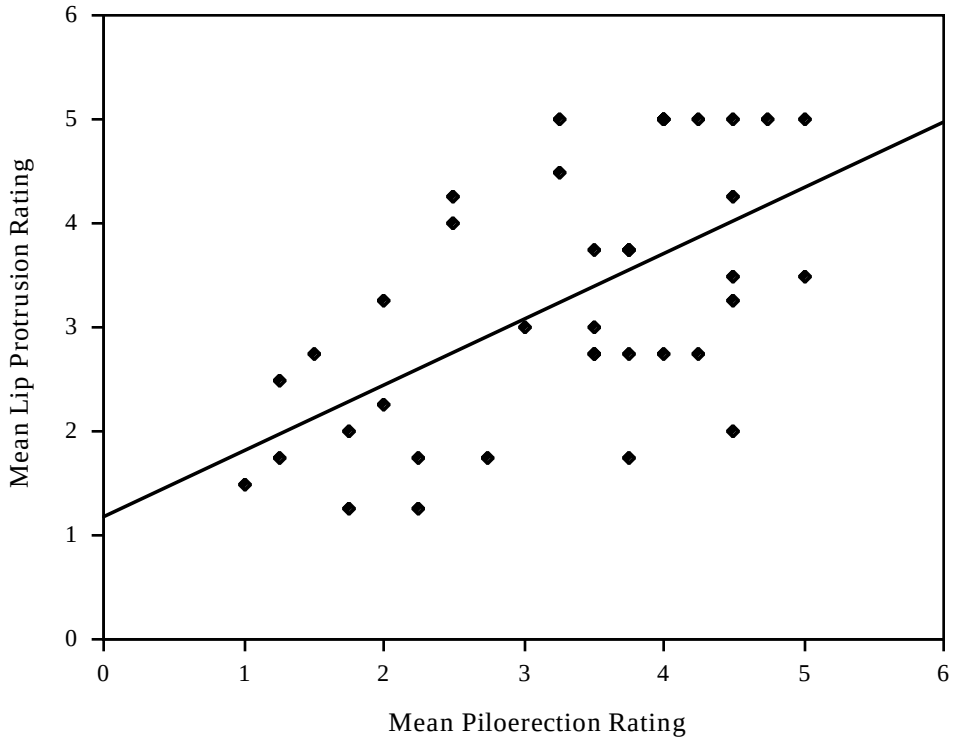


Table 11: Spearman Rank Correlation Coefficients for Piloerection Ratings vs. Lip Protrusion/Rounding Ratings for Each Rater Individually.

Rater #	Spearman rho	p value
1	.350	.0332
2	.075	.6475
3	.678	.0001
4	.529	.0013

All subjects showed a strong correlation between their judgments of piloerection and "trying to look big" (Table 12). This is unsurprising both because one of the obvious effects of piloerection would be to increase the effectiveness of such attempts, and because postural modifications such as shoulder hunching and back arching were adopted at the same time as piloerection. This raises the possibility that the piloerection ratings were "contaminated" by the raters paying attention (even subconsciously) to other postural cues. While this is certainly possible, it makes little difference to the hypothesis under test here: we are interested in the correlations between the acoustic display and an independent visual display, which allows us to assess whether the animal is attempting to increase projected size. The postural modifications, or any other aspects of the visual display, provide an equally suitable independent measure. (Indeed, using the postural ratings results in no change in the results). The only contaminant which would be significant would be a pre-existing bias on the part of the raters to link lip protrusion to piloerection (because the sound was turned off during the piloerection rating task, there was clearly no effect of sound on ratings).

Table 12: Spearman Rank Correlation Coefficients for Piloerection Ratings vs. Apparent Body Size Ratings for Each Rater Individually.

Rater #	Spearman rho	p value
1	.731	<.0001
2	.796	<.0001
3	.862	<.0001
4	.911	<.0001

Discussion

The results of this experiment indicate that when saki monkeys increase their visually-apparent size via piloerection or other cues, they also tend to protrude their lips during vocalizations. The data are thus consistent with the hypothesis that the monkeys to whom these calls are directed use vocal tract length (presumably via its acoustic correlate, formant dispersion) to assess the body size of the vocalizer. This conclusion is based on three assumptions. First, that the acoustic effect of lip protrusion is a lengthening of the vocal tract. The support for this assumption is detailed in Chapter 1 of this thesis (see also Stevens and House 1955). The second is that the manipulation of vocal tract length seen in these displays indicates a sensitivity to the affects of vocal tract length on the part of listeners, an assumption consonant with the widely shared view that animal displays evolve in keeping with the perceptual abilities of their intended audience. This assumption has broad theoretical and empirical support (see, for example, Andersson 1980 and Krebs and Dawkins 1984). Finally, this conclusion requires that piloerection is a display selected due to its effect of increasing projected body size; I now review the evidence supporting this assumption.

Piloerection and Body Size

Piloerection is a extremely widespread feature of both avian and mammalian visual displays. (In this discussion I follow Morris (1956) in terming feather erection in birds "piloerection", avoiding the more accurate but awkward term "penne-erection"). In many bird species, males use piloerection as a part of courtship displays, and a large number of species perform threat displays with the body feathers ruffled (Morris 1956). Morris (1954) reported that Zebra finches (*Poephila guttata*) fluff the plumage to inhibit attacks by dominant individuals. In many cases, certain areas of the body are covered with specialized feathers which make piloerection more dramatic, e.g., the crests of cockatoos, the ruffs of game birds, ear-tufts in owls, breast feathers in the European robin (*Erithaceus rubecula*). The vast majority of displays involving piloerection in birds take place in either agonistic or sexual contexts (Morris 1956). However, there are a few instances where this is not true: birds often erect the feathers before sleeping, and some group nesting birds have come to use piloerection as a social signal which initiates roosting. Such examples definitely are exceptions rather than rules, and Hingston (1933) summarized the data for birds thus: "All birds ruffle their feathers in anger... were I to give all my examples of feather-fluffing under anger it would mean a list of every bird whose fighting behavior I have seen recorded".

Although I have found no such sweeping statement for mammals, piloerection appears to be equally prevalent in our own class (Poduschka 1977). In voles (a mouse-sized rodent of Europe), a dominant individual preparing to attack "may assume a hunched attitude with hair erected. From the point of view of the observer the erection of the hair...makes the aggressive vole look very much larger. It is often surprising to see how small an animal really is after he has 'cooled down'" (Clarke 1956). In the acouchi *Myoprocta pratti*, a large forest rodent of South America, the hair around the flanks is erected during aggressive activities, and sometimes during courtship (Kleiman 1971), with the effect of

increasing apparent body size (Eisenberg & Kleiman 1977). In raccoons, the typical threat posture of a dominant individual involves "elevating the tail, laying back the ears, and raising the shoulder hackles", while the submissive posture assumed by a subordinate involves lowering his chin, neck and ventral body surface to the floor and retreating (Barash 1974). Similar displays, in similar contexts, occur for a range of other animals too vast to record here, including rats (Albert et al. 1992), wolves (Schenkel 1947), cattle (Antonius 1939) and many New and Old World monkey species (e.g., Moynihan 1970, Oppenheimer 1977, Gautier and Gautier 1977).

Several interesting variants of the standard mammalian piloerection display are found: the sea otter *Enhydra lutris* lacks pilomotor muscles and thus cannot piloerect, but still adopts a defensive posture, hunching the back up to look larger (Kenyon 1969). Similarly, the owl monkey *Aotus trivirgatus* does not piloerect during aggressive interactions, but instead adopts a convex arch in its spine to give a similar effect (Moynihan 1967).

A more extreme development of the pilomotor system is seen in animals with spines such as tenrecs (large insectivores restricted to Madagascar), hedgehogs, and porcupines: in these species the erection of hair and spines serves not only to make the animal appear larger, but also to ready the animal's spines to effectively repel an attack. A tenrec with its spines piloerected will move toward and threaten potential predators, swaying its head to display the erect quills (Eisenberg & Gould 1970). Interestingly, tenrecs are apparently the only species in which piloerection has an important acoustic effect: some of the quills at the base of the spine have been modified into a stridulatory organ similar to those in some insects (e.g., crickets). This organ produces a broad band acoustic signal (which includes ultrasonic components) which apparently serves both to alert predators and to maintain mother-infant contact (Eisenberg & Gould 1970). In these mammalian species with spines, the piloerector muscles appear to be highly developed: Poduschka (1977) observed that the erector muscles are so fast in the hedgehog that an attempt to photograph their action with a shutter speed of 1/125 of a second still resulted in blurred photos.

These examples (which could be amplified *ad infinitum*, see Hingston 1933) give a sense of just how widespread the use of piloerection is as a visual display. In the vast majority of cases, piloerection occurs in agonistic or sexual situations, and this has led many researchers to make the rather obvious point that piloerection serves to increase the projected size of the signaler (e.g., Clarke 1956, Eisenberg & Kleiman 1977, Moynihan 1970, Poduschka 1977). However, some researchers following the tradition of animal communication as an expression of emotion have focused on piloerection as an involuntary response to generalized autonomic activation (e.g., Gautier & Gautier 1977, Morris 1956). Although it is certainly true that piloerection involves activation of the autonomic nervous system (the arrector pili muscles receive sympathetic innervation), and I make no claims that animals are conscious of their attempts to increase body size, this observation is flawed as an explanation of the widespread use of piloerection in aggressive and sexual displays on a number of counts.

Aggressive situations, which involve high levels of sympathetic arousal for both participants, often lead to opposite piloerection displays, with the successful or dominant animal fully piloerecting and the subordinate individual completely unerected or even depressed or "sleeked". Birds have both an arrector pennae and a depressor pennae muscle, and can thus actually depress the feathers tighter against the body. (Although several researchers have described “sleeking” in mammals impressionistically (e.g., Goodall 1986, Clarke 1956), I have been unable to find any mention in the mammalian literature of a "depressor pili" muscle which would account for these observations.) In any case, however, these observations run contrary to the notion that piloerection is a side-effect of sympathetic arousal. Similarly, in isolated cases, piloerection is observed during non-aggressive contexts where, if anything, parasympathetic arousal would be expected. An example of this is the roosting response in some birds described previously (Morris 1956).

A more theoretical criticism derives from a perspective on the evolution of communication behavior popularized by Krebs and Dawkins (1984). In most situations, it is unprofitable (and unstable over

evolutionary time) to reveal detailed and accurate information about internal states to a competitor, particularly in aggressive contexts (Krebs and Dawkins 1984). An organism which "plays its cards close to its chest" will have an advantage over one which telegraphs its intentions, because it will be more able to predict and preemptively respond to future actions ("mind reading"). Thus, in many situations, combatants which reveal the details of their arousal and motivation level accurately will be selected against. Of course, an organism may benefit from exaggerating or emphasizing its arousal level in certain situations (e.g., a territory owner might profit from always seeming "angry": highly aroused and ready to fight). Krebs and Dawkins (1984) term this "manipulation". This is no help to the "piloerection as expression of internal state" hypothesis. Precisely because of the powerful selective pressure exerted on communication in contest situations, an evolutionary arms-race between "mind reading" and "manipulation" is expected which will eliminate any epiphenomenal side-effects of internal states which are perceptible to the opponent (see Chapter 4 for more on this issue). Even if piloerection originated as a simple side-effect of high arousal, the intense selection intrinsic to aggressive and mating displays would quickly take over to act on its efficacy as a signal. Thus, the most plausible explanation of piloerection in aggressive displays is the intuitively obvious one: it is an example of "manipulation" selected to increase the apparent size of the signaler.

Taking piloerection as a valid independent measure of the projected body size of a displaying individual, the data from this study are thus consistent with the hypothesis that saki monkeys, like humans, make use of vocal tract length (via formant dispersion) to estimate body size of a vocalizer. I now turn to a review of other behavioral literature which supports the idea that a perceptual relationship between vocal tract length and body size exists in other species as well.

Auditory origins of facial signals

Many visual signals, including facial gestures, are accompanied by vocalizations (Andrew 1963, Eisenberg & Kleiman 1977). In some cases the visual display will have acoustic consequences; this is particularly true of any displays involving the mouth, lips or tongue. In such cases (which include a vast number of mammalian facial displays) this raises the interesting question of which sensory domain provided the selective force behind the display. It is entirely possible, of course, that both contribute equally to the current communicative value of the display, or that both played important parts at different times during the evolution of the display. But in the case of at least some of these displays, selection in these two sensory domains would appear to act in opposing directions. When opening the mouth, a signaler reveals to the observer teeth which are dangerous weapons, making an open mouth appropriate for aggressive and threatening displays. Open-mouth displays appear to function as a powerful threat in many mammalian species, including organisms as diverse as the hippopotamus *Hippopotamus amphibius* (Walther 1977), seals and sea lions (Winn & Schneider 1977) and macaque monkeys (van Hooff 1967).

In contrast, the acoustic effect of opening the mouth is to substantially shorten the vocal tract. If the hypothesis I am advancing in this thesis applies to all (or most) mammals, vocalizations produced with an open mouth will lead to a decrease in the acoustically-apparent body size of the vocalizer. What should be expected in aggressive situations is a closed-mouth display, where the mouth opening is confined to the end of the snout and thus maximizes the vocal tract length. Thus, the acoustic and visual effects of closed- and open-mouthed displays suggest opposite predictions about the conditions under which these displays should be used.

Open mouth displays are widespread as submissive displays in at least two mammalian taxa: canids and primates. In all canids (i.e., foxes, wolves, coyotes and dogs), a "grin" involving retraction of the mouth corners is used to indicate submission (Fox & Cohen 1977). All canids also share a characteristic threat face, described by Fox & Cohen (1977: p. 734) as an "aggressive pucker or mouth-closed, lips forward expression". This is commonly seen in growling dogs (pers. obs.). In some canid species, including dogs, this threat expression gives way during intense aggression to a more open-mouthed expression where the lips are withdrawn vertically to reveal the teeth (the growl turns to a

snarl). This would appear to combine the maximal vocal tract length compatible with the visually-effective display of teeth (and a preparation to bite), as the vocal tract length is still much greater than that in the submissive grin, where the horizontal retraction of the lips (Fox & Cohen 1977: p. 735) would serve to reduce vocal tract length to its minimum.

In primates, a vast number of species share a submissive "grin" facial display (van Hooff 1967, 1972). This is true in both New World monkeys (Oppenheimer 1977) where it is variously ascribed a defensive or submissive role and is occasionally seen during play, and in Old World monkeys (van Hooff 1967, Gautier & Gautier 1977) and apes (Goodall 1986, Marler & Tenaza 1977). The human smile is also an example of this display, although its use has been extended into non-aggressive situations to denote an unthreatening or friendly attitude (see van Hooff 1972). The use of mouth corner retraction to signal submission and a lack of threat appears to be extremely widespread in primates.

There is also more limited evidence for vocal tract elongation being associated with threat in the primate order (besides the data presented in this chapter). Epple (1967) described a "protruded lips face" in the small New World tamarin *Saguinus geoffroyi* which functions as an aggressive threat. Van Hooff (1967: p.18), in his general review of primate facial displays, cited the "tense-mouth face" in which the "mouth corners are brought forward. As a result the mouth often looks like a narrow slit". This display is usually performed by the dominant animal immediately preceding an attack, and is associated with a low-pitched bark in at least some species (chimpanzees and baboons). These observations, while somewhat sparse, suggest that a variety of primate species manipulate their vocal tract lengths in behavioral situations in accordance with the acoustic predictions. Perhaps because there is no general appreciation amongst primate ethologists of the acoustic effect of lip protrusion, it is rare to find anything more than an informal mention of lip protrusion in the primate literature. However, the widespread data for lip retraction, combined with this lip protrusion data, suggest that vocal tract length is an important variable throughout the primate order, and not just for humans and saki monkeys.

There are scattered data for other taxa. Bears protrude their upper lips during aggressive interactions (Pruitt & Burghardt 1977), and my personal observations show that cats maximize vocal tract length while growling. These observations suggest that vocal tract length assessment is shared by a number of other mammal groups. There are, of course, examples which do not fit neatly into this framework. For example, the "pout face" shown by infants in a variety of primate species (including apes and Old World monkeys, van Hooff 1967) would clearly have an effect of increasing vocal tract length. It is difficult to know, however, why (or if) the “pouting” vocalizer is attempting to increase its projected body size. Examples like this probably abound, and I do not suggest that vocal tract length is the only important variable in the evolution of mixed facial-acoustic signals. However, the canid and primate data combine with these other behavioral observations to suggest that vocal tract length assessment, far from being limited to humans, is characteristic of the entire mammalian class.

Evolutionary Implications of Vocal Tract Length Assessment

In this chapter I have discussed how modifications of vocal tract length allow a vocalizer to change its acoustically-projected body size. However, vocal tract length assessment presumably evolved initially to allow listeners to make reasonably accurate judgments of body size, and only subsequently was this perceptual system harnessed by vocalizers to manipulate their projected body size. It seems likely that animals can still use the acoustic correlate of vocal tract length, formant dispersion, to provide a relatively accurate cue to body size in some situations. In humans articulatory manipulations result in a small amount of change in an individual's total vocal tract length: Fant's adult male subject was able to vary his average vocal tract length of 18 cm by about ± 1.5 cm via lip, tongue and laryngeal maneuvers (Fant 1960: p. 115). Unfortunately, no data are available on the vocal tract lengths of other organisms, but it seems very likely that animals with long snouts like baboons or dogs can modify their vocal tract length more drastically. Nonetheless, when the vocal tract is in its longest state (i.e. with mouth closed and lips fully protruded), formant dispersion will vary in a simple

and dependable fashion with head size. Thus, given a collection of animals vocalizing with protruded lips, a listener would still be able to determine the relative sizes of all the individuals.

In general, if listeners can use some independent visual or acoustic cue to determine the articulatory configuration used to produce a sound, they will be in a much better position to make use of formant frequencies as a cue to body size. Thus, the use of vocal tract length to assess body size would not only provide a selective force for formant perception, but also for an independent means of assessing articulation in a vocalizer. Both of these selective forces are clearly important for the evolution of human speech perception.

Even if vocal tract length were poorly correlated with body size (as may be the case in some birds species), it could still provide a cue to individual identity in calls produced with a known articulatory configuration. Thus we might predict that lip protrusion or some other extreme articulatory position (e.g., that used in producing the human [i] vowel) would be associated with calls revealing individual identity (i.e. the long calls of many primate species).

If large projected body size meant success in aggressive or mating behavior, the use of formant dispersion to gauge body size would result in selective pressure to increase vocal tract length (that is, to "fake" a larger body size). Such a selective pressure might have provided the initial basis for the descent of the larynx in humans. The human larynx is positioned differently in humans than in any other known mammal: it lies deep in the throat rather than at the back of the mouth (Negus 1949). This position results in a qualitatively different vocal tract which enables us to create a range of vowel sounds much greater than any other mammal (Lieberman et al. 1969, Lieberman 1984). However, evolution lacks foresight, and the increased vowel inventory of future generations could not have provided the initial selective pressure for larynx descent (moderate degrees of vocal tract lengthening will not result in the two-tube configuration required for the human vowel space). The human larynx position makes us more likely to choke while eating (Negus 1949), and choking accounts for a great number of deaths each year (about 4000 in the United States, Heimlich 1975). A relatively powerful selective force was necessary to overcome this evolutionary resistance, and allow us to achieve our current anatomy.

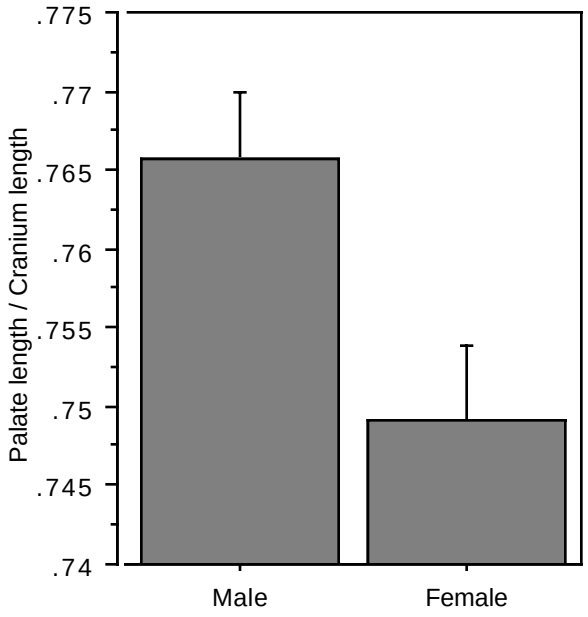
I suggest that the use of vocal tract length to gauge body size provided this initial selective force. The sexual dimorphism in human vocal tract anatomy supports this hypothesis: men (although no better able to produce speech than women) have larynges approximately 10 mm deeper in the throat than women (Senecail 1979). This dimorphism parallels a number of others which appear to differentially enhance the size difference between men and women (including the disproportionately larger larynx and resultant lower-pitched voice, and male facial hair which, like the lion's mane, may serve to increase apparent head size).

Interestingly, chimpanzees show a similar sexual dimorphism. I analyzed data from 171 *Pan troglodytes* skulls kindly given to me by physical anthropologist Robert Anemone. In both chimps and humans, the length of the palate provides a measure proportional to the total vocal tract length. Dividing this number by the total skull length (which is proportional to body size) removes the difference in palate length due to differences in total body size. A t-test comparing the palate length / cranial length index for males vs. females revealed a marked sexual dimorphism in vocal tract length (see Figure 20), over and above what would be expected due to the known dimorphism in overall body size ($t(169) = 2.617$, $p = .0097$), supporting the idea that selection has acted on male chimps to increase vocal tract length.

Thus, there is reason to believe that selection for increased vocal tract length may have provided the initial force which led ultimately to the uniquely-human vocal tract necessary to produce speech. A similar suggestion has been made by Ohala (1984), but he has missed the point that the human vocal tract has shown subsequent adaptation for speech production (Lieberman 1984), claiming instead that body size provides the only explanation for the human larynx position. I suggest instead that vocal

tract length provided the initial selective force only, serving as a preadaptation for speech-specific selection.

Figure 20: Sexual Dimorphism in Vocal Tract Length in Chimpanzees



Chapter Summary

An extensive body of ethological research supports the hypothesis that lip protrusion and mouth-corner retraction are used in mammalian displays for their acoustic effect of modifying apparent body size. The data indicate that in many aggressive or threatening contexts, in which a signaler would benefit from being perceived as larger, vocal tract length is increased. In contrast, animals in a submissive or friendly context often retract the mouth corners, decreasing vocal tract length. I suggest that the use of formant dispersion to assess body size is widespread in mammals, and has resulted in selection in our order for accurate formant frequency perception, perception of articulatory configuration during vocalizations, and vocal tract elongation. All of these selective forces are likely to have played an important role in the evolution of human speech perception and production, and in the evolution of mammalian communication in general.

Chapter 4: Vocal Tract Length and the Evolution of Human Language

In the first chapter of this thesis I showed that humans use both the fundamental frequency and the formant dispersion of a speaker's voice to assess his body size; in the second I showed that the use of vocal tract length (via formant dispersion) to gauge body size provides an explanation for the most widely-recognized form of phonetic symbolism, vowel/size symbolism. In the third chapter, I provided evidence that the judgment of size from cues for vocal tract length is not limited to humans, but also plays a role in the communication of a wide array of mammalian species. In this final chapter, I explore the issues raised by these data for two disparate fields: evolutionary theory and biological linguistics (construed broadly to include studies of speech perception and production). In the second section, I describe how animals' use of acoustic cues to assess body size promises to add empirical flesh to the bones of theoretical discussions of the evolution of communication systems. In the third, I argue that adopting an evolutionary perspective on the study of language provides interesting insights both into where language came from, and how it works today. However, because these topics presume an understanding of Darwin's (1859) theory of evolution by natural selection (in its updated modern or "neo-Darwinian" form, e.g., Futuyma 1979, Dawkins 1987, Mayr 1982), I start with a brief overview of that topic, focusing on those aspects of currently-accepted evolutionary theory which are particularly relevant to the following discussion.

The Neo-Darwinian Theory of Evolution

The idea that organisms have evolved, that is, descended with changes from common ancestors, is ancient. Our first record is from the Greek atomists (Ruse 1986), and the idea was quite well-established in the scientific community well before Darwin's writings. It was a central theme, for example, in the writings of Jean Baptiste de Lamarck, a French invertebrate zoologist, who proposed that the driving force behind evolution was the inheritance of acquired characteristics. Darwin's grandfather, Erasmus Darwin, also recognized that organisms had evolved.

There were several lines of converging evidence which pointed to evolution and common ancestry. The most striking is the fossil record, which provides evidence for a large number of life forms which have disappeared from the face of the earth, some of which are clearly precursor forms of, or even links between, extant forms. However, equally compelling evidence is provided by other sources. For example, all taxa have morphological features which are shared by, or "primitive to" most species in that group. A well-known example is the limb bones of vertebrates: there is a remarkable homology between the bones of the arm and hand in species where those limbs have been put to very different uses (e.g., in the wings of birds, the hooves of horses, the hands of moles and the flippers of seals). This strongly suggests that at some point in the distant past, forms which seem very different today shared a single less specialized ancestor, a suggestion which is supported by the fossil record. Another example comes from embryology: all vertebrates go through a characteristic set of stages during development which are similar or identical between diverse species. The fetus of a dog, at an early stage of development, is morphologically indistinguishable from that of a human fetus, and the two go through a remarkably similar set of changes as they develop.

Modern biology has added several other even more compelling lines of evidence supporting evolution. The most striking comes from sequencing of proteins and DNA: it is now clear that most proteins and the vast majority of the genome is shared between humans and other mammals, and certain key enzymes (including those responsible for maintaining and expressing the code in the DNA itself) are identical in all living forms in the plant and animal kingdoms. The ploys of Creation "scientists" (religious fundamentalists to the rest of us) notwithstanding, evolution is widely agreed to be an incontrovertible fact, and it was recognized as such by Darwin as he set out on his famous voyage.

The purpose of Darwin's "Origin of Species" was not to document evolution (which he nonetheless did quite nicely), but to explain the underlying mechanism behind it. Darwin's solution is deceptively simple: there is variation amongst individuals of a species, and due to the overproduction of young

characteristic of all known species there is a struggle for existence, in which some variants do better than others. Because some characteristics responsible for this differential survival are inherited by the offspring of the survivors, these characteristics gradually increase in the population, resulting over long periods of time in change in the characteristics of the species as a whole. I say "deceptively" simple both because this theory has been so widely and frequently misunderstood, and because Darwin's insight into the mechanism of evolution had apparently never occurred before in several thousand years of recorded history. (Interestingly, Darwin's contemporary Alfred Wallace had the same insight independent of Darwin, suggesting that it was an idea whose time had come).

Thus, the three readily-understood principles which underlie Darwin's theory are variation, a struggle for existence and heritability. Darwin did not understand the biochemical basis for either variation or heritability, and the main contribution to Darwin's theory in modern times has been the elucidation of this basis, but they were both relatively obvious to anyone with even a passing acquaintance with biology: no two animals are exactly the same, and children tend to resemble their parents more closely than some randomly-chosen member of their species. The crucial third element was provided by Malthus's observation that all known species produce vastly more offspring than could ever hope to survive. Malthus's simple calculations showed that even organisms rather slow to reproduce (such as humans or elephants) would quickly overpopulate the world if their reproduction was unchecked. Because of this fact, there is a *de facto* struggle for existence: even if organisms do not aggressively compete with one another, they will eventually have to compete for food or space once a certain population has accumulated. Darwin realized that if survival is <u>in any way</u> linked to the characteristics of the survivor (as it must in many cases be), and if these characteristics are <u>at all</u> heritable, there will be a change in the overall makeup of the population such that the heritable characteristics responsible for survival increase their proportion in the overall population.

The only significant change to Darwin's theory in the succeeding century is not a revision but an addition: we now know the underlying basis for both variation and heritability. The biochemical basis for these elements of Darwin's theory is, of course, the DNA molecule, which constitutes a genetic code specifying the structure of proteins and enzymes and regulating their expression. Observable variation in organisms derives from differences in genes, which are the functional units of heredity. This observable variation is termed "phenotypic" variation after the "phenotype", which is the actual instantiated result, after development, of a particular "genotype" or set of genes. Changes in genes come about through mutations of various sorts, which are additions, deletions or replacements of the base pairs making up the DNA.

The processes leading to mutation are random, undirected events. Although many mutations are "invisible" (they have no phenotypic effect), those which result in differences are typically fatal. This is simply an expression of the fact that organisms are highly complex interlocking systems, and making random changes in such a system (liking inserting or deleting random lines of code in a complex computer program) is more likely to do damage than good. Furthermore, genetic changes cannot profit from the experience of the organism, because the cells of the germ line (those cells which produce sperms and eggs, and thus donate their DNA to offspring) are sequestered away from somatic cells. Lamarck's idea that a giraffe which spent its life stretching to get at high leaves would have children with longer necks is thus incompatible with the biochemical nature of heredity.

<u>Attacks on, and Misconceptions about, neo-Darwinism</u>

As I mentioned above, Darwin's theory of evolution has been subjected to, and has withstood, attacks from a variety of sources. More subtly dangerous are a number of common misconceptions about the implications of the theory of natural selection. In this section I review some of these.

The most widespread misconception about Darwinism has been expressed by a bewildering variety of people. It is that "Darwin's theory attempts to explain the detailed complexity of life through a random process of evolution". For example, the eminent astronomer Hoyle (Hoyle & Wiramasinghe

1981) suggested that the idea that anything as complex as any living organism could be put together "by chance" is as likely as the idea that a storm could blow through a junkyard and assemble a Boeing 747. This idea results from an incomplete understanding of Darwin's theory: although the variation which fuels evolution is random, the selection which acts on it is anything but random, and instead is highly directed and sensitive to extremely small variations in fitness. Over a long period of time, where slight variations in a positive direction are selected for, this can result in the cumulative development of highly complex structure. Hoyle's point that the appearance of a complex structure like the eye all at once by chance is vanishingly improbable is correct, but has no bearing on the theory of natural selection.

A related misconception is that traits which have evolved must vary. Variation provides the "fuel" for natural selection, but it is quite possible for natural selection to "use up" this variation. When one particular variant of a gene, called an allele, is selected for until it becomes the only example of that gene, it is said to have reached "fixation". Although this will occur rarely in practice (if only because of continual mutation), certain crucial molecules may be so well-adapted that they go to fixation, and any mutational change leads to a performance decrement and is thus eliminated by further selection. Thus, variation in a modern population is a requirement for any future evolution of a trait, but the presence or absence of variation in a trait in a contemporary population tells us nothing about whether or not the trait evolved in the past.

Another common misconception is that evolution leads to greater and greater perfection, and that modern organisms are "better adapted" than their primitive forbears. As comforting as this assumption may be to our egos, it derives no support from the Darwinian theory of evolution, which suggests that different organisms can be equally well-adapted to their environments, with one no better suited on some ultimate scale than others. "Primitive" to an evolutionist has no pejorative connotation: as applied to a shark vs. a grouper, for example, "primitive" means that the shark has retained characteristics of an shared ancestor of sharks and groupers to a greater extent than has the grouper. Although attempts to save the idea of "progress" in evolution by recourse to the intuitive idea of complexity have been made (e.g., Dawkins 1987), the question always becomes "complexity of what?" In the comparison between fish and mammals, for example, mammals have a more complex brain, while the skin histology of most fish is much more complex than that of any mammal (Williams 1966, p 43). There is no harm in believing that the former is the more important trait (I certainly do), as long as we realize that the theory of evolution provides no justification for this belief.

A final common claim about Darwinism is that it is circular or tautological: evolution is "the survival of the fittest" and the "fittest" are "those which survived". First of all, the phrase "survival of the fittest" was coined not by Darwin, but by his contemporary Herbert Spencer (the first proponent of the idea of evolution as progress, which has been termed "Spencerism" as a result (Ruse 1986). Darwin never claimed, nor do contemporary biologists, that the fittest always survive: there is clearly a significant random and arbitrary element to survival. The point is that any non-random contribution to differential survival is enough to assure evolution. More significantly, although current evolutionary theory makes frequent use of the term "fitness" to refer to reproductive success because that is clearly the measurable aspect of an animal's quality that will influence future evolution, we can always cash out this concept of fitness in terms of the quality of design (in an engineering sense) of a particular trait for the purpose to which it is put. (Several writers have given a more detailed exposition of this argument: e.g., Caplan 1977, Dawkins 1987)

In addition to the attack from outside science of creationism, which I will not discuss here, two further challenges to Darwinism have come from within its ranks, both involving the well-known popularizer of evolution, Stephen J. Gould. Although both of these challenges really regard relatively minor "tweaks" to Darwin's theory, both have been seized upon by the popular press and critics of Darwinism as heralding the "demise of the theory of evolution" (Ruse 1986). This is due at least in part to a rather zealous and polemical tone which characterizes both of the papers involved: Eldredge

and Gould (1972), and Gould & Lewontin (1979), which has understandably been the source of some rancor amongst evolutionary biologists.

Gould's first "challenge" to neo-Darwinism was the theory of "punctuated equilibria"(Eldredge and Gould 1972, Gould & Eldredge 1977), which holds that evolution is characterized by relatively rapid periods of evolutionary change (as for example, during periods of climatic change), followed by long periods of stasis. This theory is based primarily upon the evidence of the fossil record (both Gould and Eldredge are paleontologists), which can indeed be characterized in such a way. Unfortunately, reconstructing the details of evolutionary history from the fossil record is rather like trying to understand the details of a court case in which the stenographer randomly recorded one of every thousand words: the processes and events leading to fossilization are simply too stochastic to allow us to draw convincing conclusions about the pace of evolution. Fortunately, there are many other sources of evidence about evolution, including molecular biological data which strongly support the idea of gradual evolution. In any case, neo-Darwinism makes no assumption that the pace of evolution is steady.

The central bone of contention in punctuated equilibrium debate concerns the concept of "gradualism" in evolution, which was espoused by Darwin and has been widely held ever since. Gradualism is based on the idea that mutations which result in sudden, drastic phenotypic changes ("sports" or "macromutations") will be unsuccessful, because they will destroy the carefully-achieved adaptive balance of physiological elements within the organism, and between the organism and its environment. If we imagine a species which has evolved to be near a local peak in the "adaptive landscape", evolution by mutations with gradual phenotypic effects promises to allow the species to climb the hill, while a macromutation would be the equivalent of transporting the organism to a totally different and random region, where it is extremely unlikely that the mutant trait is adaptive, and where none of the organism's other traits are adaptive. And of course, even if this new region in the adaptive landscape happened to constitute "greener pastures", who would the "hopeful monster" mate with? Although Gould (1980) cites "macromutations" as part of the argument, subsequent debate (Gould 1982) has made clear that what he really had in mind is rapid, gradual evolution, not evolution by drastic leaps. Thus, even Gould concedes that evolution is a hill-climbing process, and that a valid evolutionary sequence presupposes graded intermediate stages between forms. Thus, gradualism remains a cornerstone of modern evolutionary theory, and "punctuated equilibrium" does not, and never did, pose a real challenge to neo-Darwinism.

A more recent "attack" on neo-Darwinism came from Gould and Lewontin's (1979) critique of a point of view they labeled "adaptationism". Simply put, adaptationism is the idea that every aspect of an organism is adaptive, and that organisms constitute the best possible fit to the demands of their environment. Gould and Lewontin compared this attitude to that of the character of Dr. Pangloss, Voltaire's biting caricature of Liebniz in "Candide", who held that "this is the best of all possible worlds". Although their criticism was, in a few isolated cases, justified, the paper was basically a rhetorical attempt to convince mainstream evolutionary biologists to pay more attention to the factors which prevent evolution from leading to perfect adaptation: exaptation, phylogenetic inertia, architectural constraints, chromosomal linkage, and genetic drift. All of these factors were widely known, and widely discussed, before Gould and Lewontin's (1979) paper (the first three were discussed in detail by Darwin, who used the term "preadaptation" rather than Gould's "exaptation"). As a result, this debate was again something of a tempest in a teapot, and far from striking a mortal wound at Darwinism, constituted a small step in an ongoing debate.

In summary, the neo-Darwinian theory of evolution by natural selection is a solid well-tested empirical theory which explains the fact of evolution by virtue of three simple principles which are readily observable in everyday life. The theory holds that evolution is slow, conservative and gradual, and that due to the random nature of mutation, evolution has no foresight. However, over the time spans available since the origin of life on earth, evolution by natural selection, acting on random mutation, has been able to produce a huge number of species incredibly well-adapted to their

environments, the complexity of any one of which dwarfs that of any inorganic structures known (see Dawkins 1987 for more on this point). The theory of neo-Darwinism has withstood the attacks of a legion of critics from a wide variety of disciplines, scholarly and otherwise, and is widely accepted as the only theory which can explain the remarkable diversity and complexity of life. No non-magic alternatives have been proposed in the century since Darwin's "Origin of Species", and, given the apparent intuitive dislike that many people hold for Darwin's theory, it seems rather likely that this is not from want of trying, but because there are no adequate alternatives.

The Evolution of Communication and Honesty in Advertising

I now turn to considering some of the implications of the data presented in this thesis to an area of very active research in contemporary evolutionary theory: the evolution of communication, and in particular the evolution of aggressive displays. Although both sexes exhibit aggressive behavior in most animal species, male/male interactions are more frequently characterized by aggression, because males which can dominate other males can obtain a disproportionate share of mates. Female reproductive success is, in many species, limited by factors other than the number of males she can mate with, such as the number of eggs she can successfully produce, or the number of offspring she can successfully raise. Early ethologists observed that intraspecific aggressive encounters are often resolved through "displays": stereotyped sequences of movements and/or vocalizations which in some cases escalated into actual fighting, but more frequently resulted in resolution of the dispute without violence. These researchers (e.g., Lorenz 1966, Huxley 1966) wondered why so few actual fights occurred, and concluded that species which frequently escalated into deadly violence would rapidly be wiped out, while species which resolved disputes through less dangerous conventional displays would persist.

Unfortunately, this argument makes use of a mode of thinking termed "group selectionism", which is the idea that evolution can occur through the differential survival of groups (e.g., populations or species) of organisms, rather than through the differential survival of individuals as proposed by Darwin. Although there is nothing patently wrong with group selection (it is certainly true that some species die out, probably due to their characteristics), it is unlikely that the types of complex morphological and behavioral patterns that we typically seek to understand in organisms can be explained by group selection. This is because individual selection happens so much faster, and is so much more powerful, that it will typically "swamp" the weaker forces of group selection. Imagine a peaceful species in which all disputes were resolved conventionally. If a mutant arose which always attacked without warning, its genes would quickly spread through the population unless some other selective force, which acted on individuals, held it back. Although group selection might have the last word when the species went extinct, it could play no role in the evolutionary path taken by the species. Thus, an adequate explanation of contest resolution via conventional displays required a selective force which acted on individuals.

In the early 1970's, John Maynard Smith and his colleagues (Maynard Smith 1976, Parker 1974) provided a solution to this dilemma based on a mathematical technique borrowed from economics called game theory. In game theory, the best strategy to play depends upon what other organisms in the population are doing. Maynard Smith formalized a version of game theory which sought the "evolutionarily stable strategy" or ESS for a game, given a set of known strategies and payoffs.

The classic example is a game called "Hawks and Doves" (Krebs & Davies 1981). The Hawk strategy is simple "always fight" (and risk injury), while Doves display, and retreat if attacked (and are never injured). Members of the population play against other organisms, randomly chosen, and increase their numbers according to their success in each round of the game. In a population made up completely of Doves, a Hawk would clearly do very well, since it always wins against Doves. As a result, Hawk genes would spread rapidly through the population, and there would soon be many Hawks. Now, however, a Hawk stands a good chance of coming up against another Hawk and possibly losing and being injured (we assume for simplicity's sake that Hawks win against Hawks, and Doves

against Doves, half of the time). If we say for example that the payoff from winning is +50, the payoff for a Hawk being injured is -100, and the cost of displaying for a Dove is -10, we can construct a payoff matrix like the one below:

Attacker	Opponent	
	Hawk	Dove
Hawk	$\frac{1}{2}(50)+\frac{1}{2}(-100)=-25$	50
Dove	0	$\frac{1}{2}(50-10)+\frac{1}{2}(-10)=+15$

We can see that given the payoffs described above, a population of all Hawks would not be stable, since the average payoff is -25, and a Dove could do better with a payoff of 0. So each strategy does better if it is relatively rare, and the ESS will obviously be a mixture of the two strategies. If h is the proportion of Hawks, (and thus 1 - h is the proportion of Doves), the average payoff for a Hawk will be

$$\overline{H} = -25h + 50(1-h)$$

While for a Dove the average will be

$$\overline{D} = 0h + 15(1-h)$$

Setting these two average payoffs equal to one another, we can calculate that with the proportion of Hawks in the population equal to 7/12, the average payoff to Hawks and Doves is equal. With this makeup, the population will be in stable equilibrium. Such a population could be achieved either by having that proportion of individuals in the population always take the Hawk strategy, or by having every individual "play Hawk" seven-twelfths of the time. Of course, the proportion of hawks in an ESS is dependent upon the payoffs chosen (it can be easily shown that, as long as the cost of injury is greater than the reward of winning, there will be some Doves in an ESS). The ESS is also obviously dependent upon the definition of the strategies chosen. What this exercise demonstrates, however, is that what constitutes a "good" strategy depends on what everyone else is doing, and in many situations straightforward individual selection will act against an "always fight" strategy and for a conventional display. Game theory thus provides a demonstration that no recourse to group selection is required to explain the evolution of conventional displays.

Although the original application of game theory was to contest situations, the approach was soon applied to the more general study of communication. In an influential paper, Dawkins & Krebs (1978) argued that the "traditional ethological approach", which viewed animal displays from the point of view of information transfer, were flawed, and offered their own "cynical" view: that individual selection would lead inevitably to a situation of "manipulation", where the signals and displays performed by an actor are best thought of as attempts to control the reactor, rather than giving it information. The traditional view was based on the mistaken idea that communication evolved for the mutual benefit of actor and reactor. In general, argued Dawkins and Krebs, if any information at all is transmitted, it is likely to be false information.

In a vigorous rebuttal of this argument, Hinde (1981) argued that Dawkins and Krebs had grossly misrepresented the views of the ethological mainstream, and that few researchers thought that most communication evolved for the mutual benefit of actor and reactor. Furthermore, Hinde pointed out that there are many situations where actors do convey information, and that an observer properly attuned to that information could make use of it. In a conciliatory revision of their earlier views, Krebs and Dawkins (1984) offered a new synthesis, in which they posited that the evolution of

communication can be characterized by an interplay between "mind reading" (by observers who make use of inadvertently emitted cues) vs. "manipulation" (by actors who exploit the sensory systems of observers). Part of this theory holds that in assessment of a competitor's "resource holding potential" or RHP (Parker 1974), observers will inevitably focus on signals which are honest reflections of the organism's size or strength. It is in this context that the data from this thesis become interesting.

The game theory approach was obviously never intended to provide detailed models of animal behavior, but to provide a framework within which to analyze the evolutionary trends associated with various behavioral strategies. The problem with this approach is that, at least as simply applied, it ignores several aspects of biology which may be crucial in evolution, particularly for the evolution of social behavior. In fact, the same criticisms which have been leveled by Gould and Lewontin (1979) at the "adaptationist" stance are appropriate here. First, these simple game-theoretic models are not dynamic; they do not take account of the effects of time lags for adaptation, which may induce limit-cycles or other unusual behaviors into the system. Second, they do not take account of the effect of possible architectural constraints on signal honesty: it may be that some signals which initially provide accurate indicators of the quality or RHP of their bearer are relatively unconstrained, and thus can easily be used to "fake" higher quality, while others may be much more constrained by architectural or allometric considerations. Finally, they ignore the role of phylogenetic inertia: the fact that it may take a long time before a "mind-reading" mutant arises which has the requisite perceptual abilities to take advantage of a cue, and that once such an ability has spread it may take a long time to remove it when it is no longer accurate.

Krebs and Dawkins (1984) and many others (e.g., Gouzoules & Gouzoules 1990, Scherer 1985) innocently accept Morton's (1977) assertion that fundamental frequency will provide an "honest" (that is, difficult to fake) cue to body size. Although this idea is intuitively plausible, the available evidence suggests that this is not true, and that vocal cord mass and length (the main determinants of F_0), may be relatively easy to modify independent of body size (Chapter 1 of this thesis). The most extreme example of this is probably the hammerhead bat *Hypsignathus monstrosus*, which has a larynx which is one-third the length of its body and proportionately enlarged vocal folds. However, we need look no further than our own species for a more moderate case: at puberty, the male larynx enlarges significantly due to the influence of testosterone, resulting in a pitch difference between adult males and females which is vastly disproportionate to the size difference between sexes. Thus, although an analysis of vocal production in more species is necessary to make this prediction concrete, current data suggests that we should reevaluate the widely-accepted idea that fundamental frequency provides a difficult-to-fake or "honest" cue to body size

In mammals and perhaps anurans, vocal tract length probably provides a more dependable cue to body size, since it is typically tied directly to the size of the head, which is in turn allometrically coupled to body size (see Chapter 1). However, in birds, vocal tract length is less constrained, and greatly elongated vocal tracts are typical of the males of a number of bird species. Such differences between species allow us to address the theoretical claims raised by Krebs and Dawkins (1984) experimentally. If they are correct, and perceivers make use of only honest cues, then listeners in species where larynx size varies widely and is uncorrelated with body size (such as hammerhead bats) would be expected to ignore fundamental frequency in assessing an opponent's body size (or a possible mate's quality). Similarly, in birds like the trumpet manucode *Manucodia keraudrenii*, which has an enormously elongated vocal tract with the trachea coiled up in loops between the skin of the chest and the pectoral muscles, the theory would predict that formant frequencies would be ignored by listeners. However, both of these species have closely-related congeners which lack hypertrophied vocal organs; in these species the perceptual mechanisms are likely to be found (they must have been there primitively, or selection never would have favored hypertrophied vocal organs in the first place). Any one, or all of these predictions is easily falsifiable, and regardless of the outcome, we will have learned a great deal about the pace and mode of evolution.

In the data reported in this thesis, subjects' ratings of body size were strongly influenced by F_0, a cue which does not appear to be well-correlated with body size in humans. Fundamental is correlated with body size in toads (and perhaps some primates as well, see Hauser 1993). This suggests that there may be a substantial lag between the time when a cue becomes decorrelated (due to faking or possibly other factors) and when listeners stop using it. There are several reasons that this could occur. Even if a cue provides a relatively poor predictor for a trait such as body size or strength, it might still provide some useful information, and ignoring it might be maladaptive. While there might be strong pressure selecting for the perception of new and more "honest" cues, pressure to remove the old cue may be relatively weak, leaving the perceptual equivalent of a "living fossil". This is the result of phylogenetic inertia: evolution is not instantaneous, logical or optimal, and the selective pressures maintaining a complex interlocking system of adaptations can prevent the evolution of a new (perhaps more adaptive) mechanism. Second, it may be possible to enhance the accuracy of the original perceptual mechanism by constraining the situations in which it is employed. As mentioned in Chapter 3 for the example of lip protrusion, variability in production will not cause a problem if there is an independent way of knowing which variant is currently in use.

Another evolutionary force might underlie the retention of non-optimal perceptual mechanisms. Perhaps the mechanism responsible for F_0 estimation was originally selected for due to the F_0/body size correlation, but later became useful in other ways (e.g., in estimating the current emotional or motivational state of the vocalizer). If the influence of F_0 perception on body size judgments is not particularly costly, and the new use for emotional judgments quite useful, the old size assessment mechanism might be preserved despite its lack of current utility. This is the same idea as that behind pleiotropy, in which a gene has more than one effect on the phenotype of its bearer (most genes are pleiotropic). Pleiotropy can actually serve to maintain deleterious traits: if selection for the adaptive trait is stronger than that against the harmful hitchhiker, the gene will nonetheless increase its frequency in the population.

While none of the proposals sketched above are original, deriving from theoretical models that have been in the literature for decades, the behavioral systems involved in the assessment of body size from acoustic cues in a variety of species provide an unparalleled opportunity to test these theoretical predictions. Thus, studies of the acoustic basis for, and perceptual mechanisms underlying, animal body size judgments have much to offer theories about the evolution of animal communication systems, completely independent of any role they may play in our understanding of the evolution of human language. In the last section, I consider the possible role of vocal tract length in the evolution of human language.

Evolutionary Constraints and Innate Mechanisms

In general, evolution is an extremely slow process. It may well be literally impossible for us to comprehend just how slow it is, since the amount of time over which evolutionary change unfolds is so much greater than the amounts of time our brains evolved to grasp. The earliest traces of life on earth are three billion years old: single-celled organisms resembling bacteria. Multi-cellular life forms probably appeared about two billion years later, although our first fossils of such forms date from 700 million years ago. The first fish appeared about 500 million years ago, during the Cambrian period, while the first mammals appeared during the Triassic period, during the Age of Dinosaurs, some 200 million years ago. The divergence between humans and chimpanzees probably occurred about five million years ago, that is to say, the ancestors of our species and chimpanzees belonged to the same species at that time. Our species *Homo sapiens* is only a paltry 500,000 years old.

These huge quantities of time are necessary for the evolution of complex structures, because the "raw material" provided to natural selection is random. Evolution has no foresight: it is blind to the future. Because of the random nature of mutation, the search for new adaptive peaks is a blind groping to see what works: there are no "laws of mutation" which make certain mutations more likely just because

they would be useful at a certain time. Furthermore, selection occurs on individuals dependent on the conditions they find themselves in. While selection is a very powerful force, and very directional, it too is blind in the sense that it has no foresight: it acts not on how well future descendants of an organism might do, but on how well the organism does in the present at surviving and producing offspring. Without these unfathomably huge amounts of time for evolution to do its work, we would expect the results of its blind groping to be quite unimpressive. Consider that it took at least two billion years after the basic building block of life (the cell) appeared for multi-cellular life to evolve.

Compared to many other organisms, humans are an extremely recent invention. We can find fossil dragonflies in lumps of coal 300 million years old which are nearly indistinguishable from modern dragonfly species, and horseshoe crabs have persisted with little morphological change for more than 400 million years. In the world of mammals, bats apparently diverged from the rest of mammals about 100 million years ago (Hill 1990). The amazing thing is that, even given these huge amounts of time, the differences between species separated by gulfs of eons are so remarkably small. Bats may appear at first sight to be quite different from, say, a macaque monkey (bats fly, after all, and sleep during the day, and eat insects). But beneath these superficial differences, closer examination reveals that the similarities are much more pronounced. This is true whether one looks at gross anatomy (the structure of the skeleton), biochemical makeup (the structure of hemoglobin or myosin), behavioral interactions (mother/infant dynamics) or neuroanatomical structure (the locations of the different primary cortices). Bats are quite different from monkeys, but none of these seem to be qualitative differences. Even when we look at the one thing about bats that we know is very different from other terrestrial species, their use of echolocation to locate insects, we find surprisingly enough that the echo-delay sensitive neural mechanisms underlying this ability seem to be shared with cats and monkeys (James Simmons, pers. comm.). Given the amount of time that bats have been diverging from primates (twenty time longer than the divergence time between man and chimp), it is quite remarkable that the significant differences would be so limited. In fact, Carolus Linnaeus, the father of modern taxonomy, found bats and primates so similar that he classified them together in the same order (not knowing, of course, that their last common ancestor was 100 million years dead).

These points are quite obvious to the student of taxonomy (or any biologist interested in evolution), and they have played an important role in taxonomic and paleontological debates. However, the implications of these basic observations have rarely been considered outside of biology. It is common in several branches of the behavioral sciences (particularly psychology and linguistics) to postulate that innate mechanisms underlie behavior which is difficult to explain through learning. Because evolution is the only known process by which a behavioral pattern of any complexity would come to be genetically determined, saying that a given behavior is genetically determined is equivalent to saying that it has evolved. (Of course, very simple structures or behaviors could be the result of genetic drift or single random mutations, but innate mechanisms are typically proposed for much more complex phenomena than could be explained without natural selection). Thus, proposing that a given piece of knowledge or behavior pattern is "innate" does not solve the acquisition problem, but merely relocates it. Now the question is how did evolution lead to the acquisition of the trait in question. Unfortunately, in those cases where it is difficult to explain how a human infant (which is an extremely sophisticated learning machine) could learn something, it is much more difficult to explain how the random process of mutation acted upon by natural selection could "learn" the same thing (particularly because there is apparently no direct way to encode specific patterns of neural activity in DNA).

Of course, this is not an argument against the existence of genetically-determined patterns of behavior: no one (at least not in the field of evolutionary biology or ethology) doubts their existence. It does suggest, however, that evolutionary theory can provide constraints, sometimes rather tight constraints, on what can plausibly be considered to have evolved in a certain period of time. In particular, we should expect that complex innate mechanisms required long periods of time to evolve, and that variants of the mechanisms will exist in closely related species.

A good illustration of this point is provided by the mammalian isolation call (Newman 1992). The young of all mammals which have been studied to date possess the ability to produce a stereotypical call when separated from their mother, which is used by the mother to regain contact with her infant. This ability appears to be innate, since it can be observed in infants only a few seconds old. The isolation cries of different species differ in their acoustic structure in a number of interesting ways, and the exact behaviors eliciting isolation cries vary from species to species. Furthermore, some species such as squirrel monkeys *Saimiri sciureus* continue to use the isolation call as adults, in order to maintain contact between the individuals in a group (this may also be true of the dolphin "signature whistle": Sayigh et al. 1990) . These differences notwithstanding, the isolation call is clearly an example of an innately-determined behavior, shared by all mammals, which has undergone slight modification in different species to slightly different adaptive ends. Examples like this could be multiplied and shown to include behaviors as diverse as aggressive displays and sexual behavior: closely-related species exhibit variants of innately determined, homologous behaviors.

From this point of view, it should be clear that innate mechanisms which share no homology with mechanisms seen in other species should be rare. If we further posit that the genes underlying these same mechanisms have gone to fixation (so that the mechanisms themselves are shared by all members of the species), we should expect them to be rare indeed. This is not to say that such mechanisms could never exist, but that the slow pace of evolution leads us to expect innate mechanisms which are homologous to those of closely-related species to be much more common. Thus, a theory which suggests that a number of innate mechanisms underlie a certain behavior will be rendered implausible precisely to the degree that it posits that these innate mechanisms are species-specific to the extent that they lack homologies in other species.

A number of researchers, most prominently the linguist Noam Chomsky, have suggested that the basis for human language consists of an innately-determined "language organ". This "organ" is extremely complex and thus precisely the sort of thing that would require a long period of natural selection to evolve. However, Chomsky (1965, 1966, 1986) and others (Fodor 1983, Lenneberg 1964, Liberman & Mattingly 1985) have repeatedly suggested that a qualitative gap exists between the human "language organ" and the neural mechanisms underlying communication (or any other behavior) in all other species: that there are no homologous mechanisms in any non-human species. Furthermore, Chomsky suggests that this "language organ" is shared by all humans, implying that the genes coding for it have gone to fixation in the human species. In other words, Chomsky suggests that the mechanisms which underlie language in humans are all of exactly the sort that evolutionary theory predicts would be most rare.

A more reasonable approach to the innate basis of behavior would first determine the extent to which innate mechanisms are required to explain a certain behavior (the most obvious cases are behaviors present at birth), and then attempt to construct a preliminary description of these mechanisms. A search for homologous mechanisms in closely-related species (i.e., non-human primates) could then begin. Many of the problems in the debate over the evolution of language appear to be caused by considering "the language organ" as a unitary entity, which is either present or not. Of course, any complex organic structure can be subdivided into component parts, and language is no exception. Nothing as complex as Chomsky's "language organ" could be coded by a single gene. Some of the components of language are clearly shared with most other mammals (for instance most of the anatomy and neurophysiology underlying speech is obviously homologous to similar mechanisms in all mammals, along with many of the neural mechanisms related to categorization and learning), while some of the cognitive machinery is perhaps shared only with higher apes (for instance, the ability to construct and manipulate abstract representations of the world). It is possible that there are a few mechanisms which are indeed completely unique to humans, bearing no recognizable homology to any mechanisms present in other animals, but we can only identify these special mechanisms by a process of elimination.

The mechanisms underlying formant perception that I have discussed in this thesis obviously have more to do with speech perception (and possibly phonology) than with other aspects of language. However, to the extent that phonetic symbolism is an important force in language (and this remains to be determined), it is clearly important at the level of meaning, which certainly lies at the heart of language. In any case, it seems clear that the use of formants for body size estimation by animals provides a homology with some of the mechanisms underlying human language, and the elucidation of the similarities and differences should illuminate our understanding of the innate basis for language. The discovery of such homologous mechanisms is good news for people interested in understanding the neural and genetic basis for language, because a much broader set of experimental techniques is available for the study of animal behavior, genetics and physiology.

Of course, this is one small step in the right direction, and a much more widespread effort will be necessary before we are even close to understanding the evolution of cognition and language in man and animals. My suspicion is that the most significant difference between humans and animals has to do with the ability to create abstract, hierarchical representations of arbitrary systems which are unbound to any observable stimulus (or response). Current data suggests that humans have carried the ability much further than any other species. Nonetheless, we are, in evolutionary terms, standing on the shoulders of our predecessors, and this ability is probably very firmly rooted in less-advanced conceptual abilities of ancestral primates, and thus visible in our nearest cousins. It is tempting to suppose that human-like cognitive abilities are nascent in other primates (particularly chimps), and that a few relatively small changes (such as in total brain volume or neural connectivity) are responsible for the apparently qualitative differences between our species. If this is true, it would put us in a much better position to understand ourselves, both in terms of our phylogenetic limitations and the rewards of recent evolution.

In conclusion, in theorizing about innate mechanisms we need to balance two sides of one equation. On one hand we have the constraints of learnability theory, from which we can draw conclusions about what it is possible to learn, given a detailed statement of the input (of course, we can never be sure that we have the correct characterization of the input: the analysis could easily miss subtle cues or organizational structures which are nonetheless used by the developing child). On the other hand, we need constraints from "evolvability" theory: what can evolve, given a statement of the time frame, the architectural and phylogenetic constraints, and homologous abilities of closely-related species. Unfortunately, we are presently in a poor position to provide such a statement: for the last three decades linguists have steadfastly avoided searching for aspects of human language which may be homologous to animal communication systems, and animal researchers have not, for the most part, analyzed animal communication systems in terms of homologies with language. Fortunately, however, the situation is changing, and there is good reason to believe that great strides will be made in understanding the biology and evolution of language in the coming decades.

Prospects for the Future

Biochemists are making extraordinarily rapid progress in mapping the human genome, and by the end of this decade it is very likely that we will have completely sequenced the DNA of our own and several other species. If the rapid progress that has followed the elucidation of the complete DNA sequence of the roundworm *Caenorhabditis elegans* is any indication, we can expect the growth of our knowledge about the genetic basis of human behavior to increase precipitously as a result. However, the impact of this knowledge will be limited by our understanding of the neural basis of language. There are, after all, many unknown steps between the base pairs coding a protein and an understanding of what that protein does in an adult brain, and a simple knowledge of the base pairs will not fill this gap. We will find many differences, major and minor, between the chimpanzee genome and our own. However, an understanding of the differences and similarities between human and animal communication systems, and in particular of animal homologies to neural mechanisms known to be important to language, will enable us to know which genetic differences make a difference.

If any perceptual mechanism is innate and shared between animals and humans (and it seems likely that many, if not most, are), the formant-detecting mechanism responsible for the estimation of body size from vocal tract length is a good candidate. Body size estimation plays an important role in aggressive interactions in a wide variety of mammalian species, and may figure prominently in mate choice as well. There has thus been strong selective pressure for perception of an acoustic cue which is present in all terrestrial vertebrates, suggesting that for some 300 million years natural selection has been waiting patiently for a mutant who can perceive formant frequencies. The evidence summarized in Chapter 3, while still quite sketchy, suggests that formant perception is widespread among mammals, and that it plays an important role in their communication systems. It is widely accepted that formant frequencies are the most crucial acoustic cue for speech perception, but nothing is known at present about the neural basis for the perception of formants. I believe that an understanding of formant frequency perception in animals, in the context of their natural social behavior, promises to provide a link between the perceptual processes crucial to language and the communication systems of other animals, and thus provides an excellent opportunity to further our understanding of the evolution of, and the neural and genetic basis for, the most tantalizing facet of human behavior: language.

References

Albert, D. J., Jonik, R. H., & Walsh, M. L. (1992). Hormone-dependent aggression in male and female rats: experiential, hormonal and neural foundations. Neuroscience and Biobehavioral Reviews, 16(2), 177-192.

Andersson, M. (1980). Why are there so many threat displays? Journal of Theoretical Biology, 86, 773-781.

Andrew, R. J. (1963). The origin and evolution of the calls and facial expressions of the primates. Behaviour, 20, 1-109.

Annamalai, E. (1968). Onomatopoeic resistance to sound change in Dravidian. Studies in Indian Linguistics, 1968, 15-19.

Antonius, O. (1939). Über Symbolhandlungen und Verwandtes bei Säugetieren. Zeitschrift für Tierpsychologie, 22, 263-278.

Arak, A. (1983). Male-male competition and mate choice in anuran amphibians. In P. G. Bateson (Ed.), Mate Choice Cambridge: Cambridge University Press.

Archer, J. (1988). The behavioural biology of aggression. New York: Cambridge University Press.

Atzet, J., & Gerard, H. B. (1965). A study of phonetic symbolism among native Navajo speakers. Journal of Personality & Social Psychology, 1(5), 524-528.

Baer, T., Gore, J. C., Gracco, L. C., & Nye, R. W. (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. Journal of the Acoustical Society of America, 90(2), 799-828.

Barash, D. P. (1974). Neighbor recognition in two "solitary" carnivores: the raccoon (*Procyon lotor*) and the Red Fox (*Vulpes fulva*). Science, 185, 794-796.

Barbellion, W. N. P. (1948). The Journal of a Disappointed Man. Harmondsworth, Middx: Penguin Books.

Beehler, B. M. (1991). Birds of paradise. In D. W. Mock (Ed.), Behavior and Evolution of Birds New York: W. H. Freeman and Co.

Bellugi, U., & Klima, E. S. (1978). Two faces of sign: Iconic and abstract. Annals of the New York Academy of Sciences, 514-538.

Benade, A. H. (1990). Fundamentals of Musical Acoustics. New York: Dover Publications, Inc.

Bentley, M., & Varon, E. J. (1933). An accessory study of "phonetic symbolism". American Journal of Psychology, 45, 76-86.

Berg, J. v. d. (1958). Myoelastic-aerodynamic theory of voice production. Journal of Speech and Hearing Research, 1, 227-44.

Berg, J. v. d. (1968). Sound production in isolated human larynges. Annals of the New York Academy of Sciences, 155, 18-27.

Bickley, C., & Stevens, K. (1986). Effect of a vocal tract constriction on the glottal source: Experimental and modeling studies. Journal of Phonetics, 14, 373-382.

Birch, D., & Erickson, M. (1958). Phonetic symbolism with respect to three dimensions from the semantic differential. The Journal of General Psychology, 58, 291-297.

Bladon, R. A. W. (1977). Approaching onomatopoeia. Archivum Linguisticum, 8, 158-166.

Bolinger, D. L. (1950). Rime, assonance, and morpheme analysis. Word, 6, 116-136.

Bolinger, D. L. (1965). Forms of English. Cambridge, Massachusetts: Harvard University Press.

Bond, Z. S. (1976). Identification of vowels excerpted from neutral and nasal contexts. Journal of the Acoustical Society of America, 59(5), 1229-1232.

Brackbill, Y., & Little, K. B. (1957). Factors determining the guessing of meanings of foreign words. Journal of Abnormal and Social Psychology, 54, 312-318.

Bradbury, J. W. (1977). Lek behavior in the hammer-headed bat. Zeitschrift Tierpsychology, 45, 225-255.

Broadbent, D. E., & Ladefoged, P. (1960). Vowel judgments and adaptation level. Proceedings of the Royal Society, 151, 384-399.

Brown, R. (1958). Words and Things. Glencoe, IL: Free Press.

Brown, R., & Nuttall, R. (1959). Method in Phonetic Symbolism Experiments. Journal of Abnormal and Social Psychology, 59, 441-445.

Brown, R. W., Black, A. H., & Horowitz, A. E. (1955). Phonetic symbolism in natural languages. Journal of Abnormal and Social Psychology, 50, 388-393.

Caplan, A. (1977). Tautology, circularity and biological theory. American Naturalist, 111, 390-393.

Cheney, D. L., & Seyfarth, R. M. (1988). Assessment of meaning and detection of unreliable signals by vervet monkeys. Animal Behaviour, 36, 477-486.

Chiba, T., & Kajiyama, M. (1958). The Vowel: Its Nature and Structure. Tokyo: Phonetic Society of Japan.

Chomsky, N. (1965). Aspects of the Theory of Syntax. Cambridge, Massachusetts: MIT Press.

Chomsky, N. (1966). Language and Mind. New York: Harcourt, Brace & World.

Chomsky, N. (1986). Knowledge of Language: its Nature, Origin and Use. Praeger.

Clarke, J. R. (1956). The aggressive behaviour of the vole. Behaviour, 9, 1-23.

Clench, M. H. (1978). Tracheal elongation in birds-of-paradise. Condor, 80, 423-430.

Clutton-Brock, T. H., & Albion, S. D. (1979). The roaring of red deer and the evolution of honest advertising. Behaviour, 69, 145-170.

Cohen, J. R., Crystal, T. H., House, A. S., & Neuburg, E. P. (1980). Weighty voices and shaky evidence: a critique. Journal of the Acoustical Society of America, 68(6), 1884-1886.

Darwin, C. (1859). On the Origin of Species. Cambridge, Massachusetts: Harvard University Press.

Davies, N. B., & Halliday, T. R. (1978). Deep croaks and fighting assessment in toads. Nature, 274, 683-685.

Dawkins, R. (1987). The Blind Watchmaker. New York: W. W. Norton and Co.

Dawkins, R., & Krebs, J. R. (1978). Animal signals: information or manipulation. In J. R. Krebs & N. B. Davies (Eds.), Behavioural Ecology: An Evolutionary Approach Oxford: Blackwell Scientific Publications.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. Ecological Psychology, 1(2), 195-225.

Doke, C. M. (1927). Textbook of Zulu Grammar. Cape Town: Longmans.

Dommelen, W. A. v. (1993). Speaker height and weight identification: a re-evaluation of some old data. Journal of Phonetics, 21, 337-341.

Dugmore, S. J. (1986). Behavioural observations on a pair of captive white-faced saki monkeys (*Pithecia pithecia*). Folia primatalogica, 46, 83-90.

Dunn, H. K. (1961). Methods of measuring vowel formant bandwidths. Journal of the Acoustical Society of America, 33(12), 1737-1746.

Dunning, J. B. (1993). CRC Handbook of Avian Body Massed. London: CRC Press.

Eisenberg, J. F. (1970). The tenrecs: A study in mammalian behavior and evolution. Smithsonian Contributions to Zoology, 27, 1-137.

Eisenberg, J. F., & Kleiman, D. G. (1977). Communication in Lagomorphs and rodents. In T. A. Sebeok (Ed.), How Animals Communicate Bloomington, Indiana: Indiana University Press.

Eldredge, N., & Gould, S. J. (1972). Punctuated equilibria: an alternative to phyletic gradualism. In T. J. M. Schopf (Ed.), Models in Paleobiology (pp. 115-151). San Francisco: Freeman.

Emeneau, M. B. (1969). Onomatopoetics in the Indian language area. Language, 45, 274-299.

Epple, G. (1967). Vergleichende Untersuchungen über Sexual- und Sozialverhalten der Krallenaffen (Hapalidae). Folia Primatogica, 7, 37-65.

Fant, G. (1960). Acoustic Theory of Speech Production. The Hague: Mouton & Co.

Fant, G. (1982). The voice source: acoustic modeling. Speech Transactions Laboratory Quarterly Progress and Status Report, 4, 1-25.

Fant, G., & Ananthapadmanabha, T. V. (1982). Truncation and superposition. Speech Transactions Laboratory Quarterly Progress and Status Report, 2-3, 1-17.

Fant, G., Ishizaka, K., Lindqvist, J., & Sundberg, J. (1972). Subglottal formants. Speech Transactions Laboratory Quarterly Progress and Status Report, 1, 1-12.

Firth, J. R. (1930). Speech. London: Ernest Bonn, Ltd.

Fordyce, J. F. (1988). Studies in sound symbolism with special reference to English. Ph. D., University of California, Los Angeles.

Fox, M. W., & Cohen, J. A. (1977). Canid communication. In T. A. Sebeok (Ed.), How Animals Communicate Bloomington, Indiana: Indiana University Press.

French, P. L. (1977). Toward an explanation of phonetic symbolism. Word, 28, 305-322.

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. IEEE Transactions in Audio and Electroacoustics, AV-16(1), 73-77.

Futuyma, D. J. (1979). Evolutionary Biology. Sunderland, Massachusetts: Sinauer Associates.

Gautier, J. P., & Gautier, A. (1977). Communication in Old World monkeys. In T. A. Sebeok (Ed.), How Animals Communicate Bloomington, Indiana: Indiana University Press.

Goldstein, U. G. (1980). An articulatory model for the vocal tracts of growing children. D. Sc., Massachusetts Institute of Technology.

Goodall, J. (1986). The Chimpanzees of Gombe: Patterns of Behavior. Cambridge, Massachusetts: Harvard University Press.

Gould, S. J. (1980). Is a new and general theory of evolution emerging. Paleobiology, 6, 119-130.

Gould, S. J. (1982). Punctuated equilibrium - a different way of seeing. In J. Cherfas (Ed.), Darwin Up to Date (pp. 119-130). London: IPC Magazines.

Gould, S. J., & Eldredge, N. (1977). Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology, 3, 115-151.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society, B, 205, 581-598.

Gouzoules, H., & Gouzoules, S. (1990). Body size effects on the acoustic structure of pigtail macaque (Macaca nemestrina) screams. Ethology, 85, 324-334.

Greenewalt, C. H. (1968). Bird Song: Acoustics and Physiology. Washington: Smithsonian Institution Press.

Hauser, M. D. (1989). Ontogenetic changes in the comprehension and production of vervet monkey (Cercopithecus aethiops) vocalizations. Journal of Comparative Psychology, 103, 149-158.

Hauser, M. D. (1993). The evolution of nonhuman primate vocalizations: effects of phylogeny, body weight and social context. American Naturalist, 142(3), 528-542.

Heimlich, H. J. (1975). A life-saving maneuver to prevent food-choking. Journal of the American Medical Association, 234(4), 398-401.

Henke, W. L. (1966). Dynamic articulatory model of speech production using computer simulation. Ph.D., Massachusetts Institute of Technology.

Hill, J. E. (1990). Bats. In E. Gould & G. McKay (Eds.), Encyclopedia of Animals:Mammals New York: W. H. Smith.

Hinde, R. A. (1981). Animal signals: ethological and games-theory approaches are not incompatible. Animal Behaviour, 29, 535-542.

Hingston, R. W. G. (1933). Animal colour and adornment. London: Unknown Publisher.

Hooff, J. A. R. A. M. v. (1967). The facial displays of the Catarrhine monkeys and apes. In D. Morris (Ed.), Primate Ethology London: Weidenfeld and Nicolson.

Hooff, J. A. R. A. M. v. (1972). A comparative approach to the phylogeny of laughter and smiling. In R. Hinde (Ed.), Nonverbal Communication New York: Cambridge University Press.

Householder, F. W. (1946). On the problem of sound and meaning, An English phonestheme. Word, 2, 83-84.

Hoyle, F., & Wickramasinghe, N. C. (1981). Evolution from Space. London: J. M. Dent.

Huang, Y.-H., Pratoomraj, S., & Johnson, R. C. (1969). Universal magnitude symbolism. Journal of Verbal Learning and Verbal Behavior, 8(1), 155-156.

Huxley, J. (1966). A discussion of ritualization of behaviour in animals and man. Philosophical Transactions of the Royal Society of London, 251, 249-271.

Jakobson, R. (1978). Six Lectures on Sound and Meaning. Cambridge, Mass: MIT Press.

Jakobson, R., & Waugh, L. (1979). The Sound Shape of Language. Bloomington, IN: Indiana University Press.

Jespersen, O. (1922). Language: Its Nature, Development and Origin. New York: W. W. Norton & Co.

Jespersen, O. (1933). Symbolic value of the vowel I. In O. Jespersen (Ed.), Linguistica (pp. 283-303). Copenhagen: Levin and Munksgaard.

Johnson, R. C. (1967). Magnitude Symbolism of English Words. Journal of Verbal Learning and Verbal Behavior, 6, 508-511.

Johnson, R. C., Suzuki, N. S., & Olds, W. K. (1964). Phonetic symbolism in an artificial language. Journal of Abnormal and Social Psychology, 69(2), 233-236.

Just, M. A., & Carpenter, P. A. (1987). The Psychology of Reading and Language Comprehension. Boston: Allyn and Bacon.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. Psychological Review, 99(2), 349-364.

Kenyon, K. W. (1969). The sea otter in the eastern Pacific ocean. North American Fauna, 68(1-352).

Kingdon, J. (1974). East African Mammals: Vol II Part A: Insectivores and Bats. New York: Academic Press.

Klank, L. J. K., Huang, Y.-H., & Johnson, R. C. (1971). Determinants of success in matching word pairs in tests of phonetic symbolism. Journal of Verbal Learning & Verbal Behavior, 10, 140-148.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America, 87(2), 820-857.

Kleiman, D. G. (1971). The courtship and copulatory behaviour of the green acouchi, *Myoprocta pratti*. Zeitschrift für Tierpsychologie, 29, 259-78.

Krebs, J. R., & Davies, N. B. (1981). An Introduction to Behavioural Ecology. Sunderland, Massachusetts: Sinauer Associates.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind reading and deception. In J. R. Krebs & N. B. Davies (Eds.), Behavioural Ecology Sunderland, Massachusetts: Sinauer Associates.

Künzel, H. J. (1989). How well does average fundamental frequency correlate with speaker height and weight? Phonetica, 46, 117-125.

Ladd, D. R. (1978). The Structure of Intonational Meaning. Bloomington: University of Indiana Press.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. Journal of the Acoustical Society of America, 29(1), 98-104.

Ladich, F. (1990). Vocalization during agonistic behaviour in *Cottus gobio* L. (Cottidea): an acoustic threat display. Ethology, 84, 193-201.

Ladich, F., Brittinger, W., & Kratochvil, H. (1992). Significance of agonistic vocalization in the croaking gourami (*Trichopsis vittatus*, Teleostei). Ethology, 90, 307-314.

Lande, R. (1980). Sexual dimorphism, sexual selection, and adaptation in polygenic characters. Evolution, 34, 292-305.

Lass, N. J., Beverly, A. S., Nicosia, D. K., & Simpson, L. A. (1978). An investigation of speaker height and weight identification by means of direct estimation. Journal of Phonetics, 6, 69-76.

Lass, N. J., & Brown, W. S. (1978). Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies. Journal of the Acoustical Society of America, 63(4), 1218-1220.

Lass, N. J., & Davis, M. (1976). An investigation of speaker height and weight identification. Journal of the Acoustical Society of America, 60, 700-703.

Lass, N. J., Phillips, J. K., & Bruchey, C. A. (1980). The effect of filtered speech on speaker height and weight identification. Journal of Phonetics, 8, 91-100.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74(6), 431-461.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.

Liberman, M., & Prince, A. S. (1977). On stress and linguistic rhythm. Linguistic Inquiry, 8, 249-336.

Lieberman, P. (1968). Primate vocalization and human linguistic ability. Journal of the Acoustical Society of America, 44(6), 1574-1584.

Lieberman, P. (1984). The Biology and Evolution of Language. Cambridge, Mass.: Harvard University Press.

Lieberman, P. (1991). Uniquely Human: The Evolution of Speech, Thought and Selfless Behavior. Cambridge, Mass.: Harvard University Press.

Lieberman, P., & Blumstein, S. E. (1988). Speech physiology, speech perception, and acoustic phonetics. New York: Cambridge University Press.

Lieberman, P. H., Klatt, D. H., & Wilson, W. H. (1969). Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. Science, 164, 1185-1187.

Lorenz, K. (1966). Evolution of ritualization in the biological and cultural spheres. Philosophical Transactions of the Royal Society of London, 251, 273-284.

Maddieson, I. (1984). Patterns of sounds. Cambridge: Cambridge University Press.

Malkiel, Y. (1990). Diachronic problems in phonosymbolism. Amsterdam: John Benjamins.

Maltzman, I., Morrisett, L., & Brooks, L. O. (1956). An investigation of phonetic symbolism. Journal of Abnormal & Social Psychology, 53, 249-251.

Marchand, H. (1959). Phonetic symbolism in English word-formation. Indogermanische Forschungen, 64, 146-277.

Markel, J. D., & Gray, A. H. (1976). Linear Prediction of Speech. New York: Springer Verlag.

Markel, N. N., & Hamp, E. P. (1960). Connotative meanings of certain phoneme sequences. Studies in Linguistics, 15(3), 47-61.

Marler, P. (1968). Visual signals. In T. A. Sebeok (Ed.), Animal Communication: Techniques of Study and Results of Research Bloomington, Indiana: Indiana University Press.

Marler, P., & Tenaza, R. (1977). Signaling behavior of apes with special reference to vocalization. In T. A. Sebeok (Ed.), How Animals Communicate Bloomington, Indiana: Indiana University Press.

Maynard-Smith, J., & Parker, G. A. (1976). The logic of asymmetric contests. Animal Behaviour, 24, 159-175.

Mayr, E. (1982). The Growth of Biological Thought: Diversity, Evolution and Inheritance. Cambridge, Massachusetts: Harvard University Press.

McClelland, B., & Wilcyzynski, W. (1989). Release call characteristics of male and female *Rana pipiens*. Copeia, 1989, 1045-1049.

McComb, K. E. (1991). Female choice for high roaring rates in red deer, *Cervus elaphus*. Animal Behaviour, 41, 79-88.

Miller, G. A., & Gildea, P. M. (1987). How children learn words. Scientific American, 257(Sept), 94-99.

Miron, M. S. (1961). A cross-linguistic investigation of phonetic symbolism. Journal of Abnormal and Social Psychology, 62, 623-630.

Morris, D. (1954). The reproductive behaviour of the Zebra Finch (*Poephila guttata*), with special reference to pseudofemale behaviour and displacement activities. Behaviour, 6, 271-322.

Morris, D. J. (1956). The feather postures of birds and the problem of the origin of social signals. Behaviour, 9, 75-114.

Morse, P. M. (1981). Vibration and Sound. Woodbury, New York: Acoustical Society of America.

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. American Naturalist, 111(855-869).

Moynihan, M. (1967). Comparative aspects of communication in New World primates. In D. Morris (Ed.), Primate Ethology London: Weidenfeld and Nicolson.

Moynihan, M. (1970). Some behavior patterns of Platyrrhine monkeys II. *Saguinus geoffroyi* and some other tamarins. Smithsonian Contributions to Zoology, 28, 1-77.

Müller, J. (1848). The physiology of the senses, voice and muscular motion with mental faculties (W. Baly, translator). London: Walton and Maberly.

Myrberg, A. A., Ha, S. J., & Shamblott, M. J. (1993). The sounds of bicolor damselfish (Pomacentrus partitus): Predictors of body size and a spectral basis for individual recognition and assessment. Journal of the Acoustical Society of America, 94(6), 3067-3070.

Myrberg, A. A., Mohler, M., & Catala, J. D. (1986). Sound production by males of a coral reef fish (Pomacentrus partitus): its significance to females. Animal Behaviour, 34, 913-923.

Myrberg, A. A., & Riggio, R. J. (1985). Acoustically-mediated individual recognition by a coral reef fish (*Pomacentrus partitus*). Animal Behaviour, 33, 411-416.

Nearey, T. (1978). Phonetic Features for Vowels. Bloomington: Indiana University Linguistics Club.

Negus, V. E. (1949). The Comparative Anatomy and Physiology of the Larynx. London: William Heineman Medical Books.

Newman, J. D. (1992). The primate isolation call and the evolution and physiological control of human speech. In J. Wind, B. A. Chiarelli, B. Bichakjian, & A. Nocentini (Eds.), Language Origins: A Multidisciplinary Approach Dordrecht: Kluwer Academic.

Newman, S. S. (1933). Further experiments in phonetic symbolism. American Journal of Psychology, 45, 53-75.

Nord, L., Ananthapadmanabha, T. V., & Fant, G. (1986). Signal analysis and perceptual tests of vowel responses with an interactive source-filter model. Journal of Phonetics, 14, 401-404.

Nowicki, S. (1987). Vocal tract resonances in oscine bird sounds production: evidence from birdsongs in a helium atmosphere. Nature, 325, 53-55.

O'Boyle, M. W., Miller, D. A., & Rahmani, F. (1987). Sound-meaning relationships in speakers of Urdu and English: Evidence for a cross-cultural phonetic symbolism. Journal of Psycholinguistic Research, 16(3), 273-288.

Ohala, J. J. (1980). ABSTRACT: The acoustic origin of the smile. Journal of the Acoustical Society of America, Suppl.1, 68, S33.

Ohala, J. J. (1983). Cross-language use of pitch: an ethological view. Phonetica, 40, 1-18.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. Phonetica, 41, 1-16.

Olivera, J. M. S., Lima, M. G., Bonvincino, C., Ayres, J. M., & Fleagle, J. G. (1985). Preliminary notes on the ecology and behavior of the Guianan saki (*Pithecia pithecia*, Linnaeus 1766; Cebidae, Primate). Acta Amazonica, 15, 249-263.

Oppenheimer, J. R. (1977). Communication in New World monkeys. In T. A. Sebeok (Ed.), How Animals Communicate Bloomington, Indiana: Indiana University Press.

Orr, J. (1944). On some sound values in English. British Journal of Psychology, 35, 1-8.

Osgood, C. E., & Suci, G. J. (1955). Factor analysis of meaning. Journal of Experimental Psychology, 50, 325-338.

Owren, M. J., & Bernacki, R. (1988). The acoustic features of vervet monkey (Cercopithecus aethiops) alarm calls. Journal of the Acoustical Society of America, 83, 1927-1935.

Paget, R. (1930). Human Speech. London: Kegan Paul, Trench, Trubner and Co.

Parker, G. A. (1974). Assessment strategy and the evolution of fighting behavior. Journal of Theoretical Biology, 47, 223-243.

Peters, R. H. (1983). The ecological implications of body size. New York: Cambridge University Press.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of vowels. Journal of the Acoustical Society of America, 24(2), 175-184.

Poduschka, W. (1977). Insectivore Communication. In T. A. Sebeok (Ed.), How Animals Communicate Bloomington, Indiana: Indiana University Press.

Pruitt, C. H., & Burghardt, G. M. (1977). Communication in terrestrial carnivores: Mustelidae, Procyonidae and Ursidae. In T. A. Sebeok (Ed.), How Animals Communicate Bloomington, Indiana: Indiana University Press.

Reichard, G. A., Jakobson, R., & Werth, E. (1949). Language and synesthesia. Word, 5, 224-233.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech Perception without traditional speech cues. Science, 212, 947-950.

Roberts, T. S. (1880). The convolution of the trachea in the sandhill and whooping cranes. American Naturalist, 14, 108-114.

Rothenberg, M. (1985). Source-tract acoustic interactions in breathy voice. In I. R. Titze & R. C. Scherer (Eds.), Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control Denver: Denver Center for the Performing Arts.

Ruhlen, M. (1991). A guide to the World's Languages. Volume 1: Classification. Stanford, California: Stanford University Press.

Ruse, M. (1986). Taking Darwin Seriously. New York: Basil Blackwell.

Ryalls, J. H., & Lieberman, P. (1982). Fundamental frequency and vowel perception. Journal of the Acoustical Society of America, 72(5), 1631-1634.

Ryan, M. J. (1988). Constraints and patterns in the evolution of anuran acoustic communication. In B. Fritzch, M. J. Ryan, W. Wilczynski, T. E. Hetherington, & W. Walkowiak (Eds.), The Evolution of the Amphibian Auditory System New York: Wiley.

Samuels, M. L. (1972). Linguistic Evolution with special reference to English. Cambridge: Cambridge University Press.

Sapir, E. (1929). A study in phonetic symbolism. Journal of Experimental Psychology, 12, 225-239.

Saussure, F. d. (1916). Course in General Linguistics (Wade Baskin, Trans.). New York: McGraw-Hill.

Sayigh, L. S., Tyack, P. L., Wells, R. S., & Scott, M. D. (1990). Signature whistles of free-ranging bottlenose dolphins *Tursiops truncatus*: stability and mother-offspring comparisons. Behavioral Ecology & Sociobiology, 26, 247-260.

Schenkel, R. (1947). Expression studies of wolves. Behaviour, 1, 81-129.

Scherer, K. R. (1981). Vocal indicators of stress. In J. Darby (Ed.), Speech evaluation in psychiatry New York: Grune & Stratton.

Scherer, K. R. (1985). Vocal affect signaling: a comparative approach. Advances in the Study of Behavior, 15, 189-244.

Schmidt-Nielsen, K. (1984). Scaling: Why is animal size so important? New York: Cambridge University Press.

Schneider, R., Kuhn, H.-J., & Kelemen, G. (1967). Der Larynx des männlichen *Hypsignathus monstrosus* Allen, 1861 (Pteropodidae, Megachiroptera, Mammalia). Zeitschrift für wissenschaftliche Zoologie, 175, 1-53.

Schouten, J. F. (1940). The perception of pitch. Philips Technical Review, 5, 286-294.

Schutte, H. K., & Miller, D. G. (1986). The effect of F_0/F1 coincidence in soprano high notes on pressure at the glottis. Journal of Phonetics, 14, 385-392.

Senecail, B. (1979). L'Os Hyoide: Introduction Anatomique a l'Etude de certains Mecanismes de la Phonation. Paris: Faculté de Médicine de Paris.

Sherman, D. (1975). Noun-verb stress alternation: An example of lexical diffusion of sound change in English. Linguistics, 159, 43-71.

Slobin, D. I. (1968). Antonymic phonetic symbolism in three natural languages. Journal of Personality and Social Psychology, 10(3), 301-305.

Smith, B. L., & McLean-Muse, A. (1987). Effects of rate and bite block manipulations on kinematic characteristics of children's speech. Journal of the Acoustical Society of America, 81(3), 747-754.

Sommers, M. S., Moody, D. B., Prosen, C. A., & Stebbins, W. C. (1992). Formant frequency discrimination by Japanese macaques (*Macaca fuscata*). Journal of the Acoustical Society of America, 91(6), 3499-3510.

Stebbins, W. C. (1983). The Acoustic Sense of Animals. Cambridge, Mass.: Harvard University Press.

Stevens, K. N., & House, A. S. (1955). Development of a quantitative description of vowel articulation. Journal of the Acoustical Society of America, 27(3), 484-493.

Sundberg, J. (1975). Formant technique in a professional female singer. Acustica, 32, 89-96.

Sundberg, J. (1987). The Science of the Singing Voice. Dekalb, Illinois: Northern Illinois University Press.

Sundberg, J. (1991). The Science of Musical Sounds. New York: Academic Press, Inc.

Swadesh, M. (1971). The Origin and Diversification of Language. Chicago: Aldine-Atherton.

Tarte, R. D. (1974). Phonetic symbolism in adult native speakers of Czech. Language and Speech, 17, 87-94.

Tarte, R. D. (1982). The relationship between monosyllables and pure tones: An investigation of phonetic symbolism. Journal of Verbal Learning and Verbal Behavior, 21, 352-360.

Tarte, R. D., & Barritt, L. S. (1971). Phonetic symbolism in adult native speakers of English: three studies. Language and Speech, 14, 158-168.

Taylor, I. K. (1963). Phonetic symbolism re-examined. Psychological Bulletin, 60(2), 200-209.

Taylor, I. K., & Taylor, M. M. (1962). Phonetic symbolism in four unrelated languages. Canadian Journal of Psychology, 16(4), 344-356.

Taylor, I. K., & Taylor, M. M. (1965). Another look at phonetic symbolism. Psychological Bulletin, 64(6), 413-427.

Tolman, A. H. (1887). The laws of tone-color in the English language. Andover Review, 7, 326-337.

Ultan, R. (1984). Size-sound symbolism. In J. Greenberg (Ed.), Universals of Human Language (pp. 525-568). Stanford: Stanford University Press.

Walkowiak, W. (1988). Neuroethology of anuran call recognition. In B. Fritzch, M. J. Ryan, W. Wilczynski, T. E. Hetherington, & W. Walkowiak (Eds.), The Evolution of the Amphibian Auditory System New York: Wiley.

Walther, F. R. (1977). Artiodactyla. In T. A. Sebeok (Ed.), How Animals Communicate Bloomington, Indiana: Indiana University Press.

Weiss, J. H. (1963). Role of "meaningfulness" versus meaning dimensions in guessing the meanings of foreign words. Journal of Abnormal & Social Psychology, 66(6), 541-546.

Weiss, J. H. (1964). Phonetic symbolism re-examined. Psychological Bulletin, 61, 454-458.

Wescott, R. W. (1987). Holesthemes or phonesthemes twice over. General Linguistics, 27(2), 67-72.

Wilczynski, W., Keddy-Hector, A. C., & Ryan, M. J. (1992). Call patterns and basilar papilla tuning in cricket frogs. I. Differences among populations and between sexes. Brain Behavior & Evolution, 1992, 229-237.

Williams, C. E., & Stevens, K. N. (1969). On determining the emotional state of pilots during flight: An exploratory study. Aerospace Medicine, 40, 1369-1372.

Williams, G. C. (1966). Adaptation and Natural Selection: a Critique of some Current Evolutionary Thought. Princeton, New Jersey: Princeton University Press.

Wilson, E. O. (1972). Animal communication. Scientific American, 227, 52-60.

Winn, H. E., & Schneider, J. (1977). Communication in Sirenians, sea otters and pinnipeds. In T. A. Sebeok (Ed.), How Animals Communicate Bloomington, Indiana: Indiana University Press.

Woodworth, N. L. (1991). Sound symbolism in proximal and distal forms. Linguistics, 29, 273-299.

Zahavi, P. (1977). Reliability in communication systems and the evolution of altruism. In B. Stonehouse & C. Perrins (Eds.), Evolutionary Ecology London: Macmillan Press.

Zakon, H., & Wilczynski, W. (1988). The physiology of the anuran eighth nerve. In B. Fritzch, M. J. Ryan, W. Wilczynski, T. E. Hetherington, & W. Walkowiak (Eds.), The Evolution of the Amphibian Auditory System New York: Wiley.

Appendix 1: Phonetic symbolism for size in the world's languages

Because most of the data pertaining to phonetic symbolism comes from English or related (i.e. Indo-European) languages, I performed a study which extends these results, in a simple and preliminary way, to all of the world's languages. Using the recent language classification system of Ruhlen (1987), I chose one example language from each of the 16 major language phyla (omitting Indo-European because of the ample evidence already reviewed showing sound symbolism in this language group). The criterion for language selection was simple: I went through the list of "better-known languages" belonging to each of Ruhlen's language phyla (Table 8.2, p. 286-289) in the order listed, choosing the first one documented in the Brown University library (in either a dictionary or grammar). Because neither Ruhlen's listings or the volumes in Brown's library were selected with regards to phonetic symbolism, this criterion provided an unbiased sample adequate for a preliminary investigation.

The languages chosen using this selection criterion were:

1. KHOISAN : Nama (= Hottentot) - S. Africa
2. NIGER-KORDOFANIAN: Fulani - C. Africa
3. NILO-SAHARAN: Kanuri - C. Africa
4. AFRO-ASIATIC: Hausa - C. Africa
5. CAUCASIAN: Abkhaz - Georgia
6. URALIC: Hungarian - Europe
7. ALTAIC: Turkish - Asia minor
8. ESKIMO-ALEUT: Eskimo - Alaska and Siberia
9. ELAMO-DRAVIDIAN: Kannada - S. India
10. SINO-TIBETAN: Mandarin - China
11. AUSTRIC: Vietnamese - SE Asia
12. INDO-PACIFIC: Telefol - Papua New Guinea
13. AUSTRALIAN: Guugu Yimidhir - NE Australia
14. NA-DENE: Navaho - N. America
15. AMERIND: Blackfoot - N. America
16. UNAFFILIATED: Basque - N. Iberia

No language guide from the CHUKCHI-KAMCHATKAN phylum (a group of five languages used by about 20,000 speakers in northeast Siberia) could be obtained, so this group was left out of the analysis. Basque was included to represent the residue of unaffiliated languages remaining from Ruhlen's (1987) classification. No pidgin or creole languages were used, since the well-known examples (Tok Pisin, Island Creole) derive from Indo-European base vocabularies, and could be "contaminated" by size-symbolism in the parent vocabulary.

Having obtained dictionaries, grammars or other sources of vocabulary for these languages, I then looked up the equivalents of the following word pairs: big/small, huge/tiny, large/little. The only word pair for which I consistently found definitions was big/small (if listings existed for large and little, they were with few exceptions the same as those for big and small). Tiny and huge were found in only 7 and 8 languages, respectively. The analysis below will thus consider only the translations for big and small.

In order to normalize for differences in word length and vowel distribution between the different languages (e.g., Eskimo "ongarurum" and Mandarin "da" both mean "big") my analysis proceeded on a language-by-language basis, comparing the word for "small" with the word for "large". I first removed any vowels which were paired in the two words on a phoneme-by-phoneme basis. For example, if one word had two /a/s and the other just one, I removed one /a/ from each word, leaving a "residue" of /a/

in the first word. Consonants were ignored. Then I examined the residues to search for evidence of a coherence between vowel and size. If sound symbolism plays a role in this corpus, the residues associated with words meaning large should have a preponderance of vowels produced with long vocal tracts, while words meaning "small" should have residues with vowels produced with short vocal tracts. The results are summarized in Table 1, below:

Table 1: Phonetic Symbolism for Size in the World's Languages

				"residue"		
	English	**big**	**small**	**big**	**small**	**Support?**
1	"Bushman"	dzuia	dema	u,i	e	?
2	Fulani	mawdo	petel	a,o	e	YES
3	Kanuri	kura	gana	u	a	YES
4	Hausa	baba	karami	-	i	YES
5	Abkhaz	adew	axece	-	e	?
6	Hungarian	nagy	kis	a	i	YES
7	Turkish	büyük	küçük	-	-	?
8	Eskimo	ongarurum	mikero	a,u	i,e	YES
9	Kannada	dodda	cikka	o	i	YES
10	Mandarin	da	siao	-	i,o	?
11	Vietnamese	to lon	be nho	o	e	YES
12	Telefol	afalik	katip	a	-	?
13	Guugu Yimidhir	warrga	bidha	a	i	YES
14	Navaho	coh	yazi	o	a,i	YES
15	Blackfoot	omahk	inak	o	i	YES
16	Basque	haundi	txiki	a,u	i	YES

There are NO examples in this corpus which contradict the hypothesis that vowels produced with shortened vocal tracts (typically /i/ and /e/) should be associated with the meaning "small", while vowels produced with longer vocal tracts (/a/ < /o/ < /u/) should be associated with the meaning "large". Thus, all eleven of the languages which have anything clear to say about the hypothesis are in accord with its predictions. This is a statistically significant bias (sign test, binomial distribution with p = .50, N = 11: p = .0005). (The *a priori* probability of confirmation is assumed to be 0.5, because, in the absence of any sound symbolism, any biasing factors such as uneven vowel distributions should affect both words equally). This data thus provides convincing support to the hypothesis that sound symbolism for size is a language universal.

Indeed, examination of the rest of the corpus suggests that sound symbolism is present in even more cases. For example, Mandarin "da" (big) and "siao" (small) was categorized as neither supporting nor contradicting the hypothesis, because the residue left after /a/ deletion, "sio", contains both a large and small vowel. However, examination of the Mandarin words "ju" (huge) and "wei" (tiny), or "gau" (tall) and "ai" (short) suggests that sound symbolism may still be present in other items in the Mandarin vocabulary. (A similar pattern is seen in Turkish, which was also scored as non-conclusive: "kocaman" (huge) vs. "minimini" (tiny)).

Of course, there are several major limitations to this study. The most important is that it is based not on acoustic analysis of the languages but a simple phonetic transcription (sometimes of doubtful

accuracy in the case of the older sources). Furthermore, the language sample, although very broad, is still numerically small. I would certainly expect there to be at least some examples of languages which directly contradict the prediction tested here. Nonetheless, these preliminary data provide unequivocal support for the hypothesis that phonetic symbolism for size may be, broadly speaking, a linguistic universal.

References:

Aulestia, G., & White, L. (1990). English-Basque dictionary . Reno: University of Nevada Press.

Eguchi, P. K. (1986). An English-Fulfulde dictionary . Tokyo, Japan: Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA).

Fahir, H. C. ., & Hony, A. D. (1992). The Oxford Turkish dictionary. ,

Frances, D. (1956). A Bushman Dictionary . New Haven: American Oriental Society.

Frantz, D. G., & Russell, N. J. (1989). Blackfoot dictionary of stems, roots and affixes . Toronto: University of Toronto Press.

Haile, B. (1951). A stem vocabulary of the Navaho language . St. Michaels, Arizona: St. Michaels Press.

Haviland, J. (1979). Guugu Yimidhirr. In R. M. Dixon (Ed.), Handbook of Australian languages. Canberra: Australian National University Press.

Healy, P., & Healey, A. (1977). Telefol dictionary . Canberra: Dept. of Linguistics, Research School of Pacific Studies, Australian National University.

Hewitt, B. G. (1979). Abkhaz . Amsterdam: North Holland Pub. Co.

Liu, Z. (1993). (pers comm) Mandarin native speaker informant. ,

Lukas, J. (1937). A study of the Kanuri language, grammar and vocabulary . New York: Oxford University Press.

Newman, R. M. (1990). An English-Hausa dictionary . New Haven: Yale University Press.

Nguyen-dinh-Hoa. (1966). Vietnamese-English dictionary . Rutland, Vt.: C.E. Tuttle Co.

Orszagh, L. (1990). A concise English-Hungarian dictionary. New York: Oxford University Press.

Ruhlen, M. (1991). A guide to the World's Languages. Volume 1: Classification. Stanford, California: Stanford University Press.

Sridhar, S. R. (1990). Kannada . London: Routledge.

Wells, R., & Kelly, J. W. (1890). English-Eskimo and Eskimo-English vocabularies . Washington D.C.: Government. Printing Office.