Human Speech: A Restricted Use of the Mammalian Larynx

*,†Ingo R. Titze, *Salt Lake City, Utah, and †Iowa City, Iowa*

Summary: Purpose. Speech has been hailed as unique to human evolution. Although the inventory of distinct sounds producible with vocal tract articulators is a great advantage in human oral communication, it is argued here that the larynx as a sound source in speech is limited in its range and capability because a low fundamental frequency is ideal for phonemic intelligibility and source-filter independence.

Method. Four existing data sets were combined to make an argument regarding exclusive use of the larynx for speech: (1) range of fundamental frequency, (2) laryngeal muscle activation, (3) vocal fold length in relation to sarcomere length of the major laryngeal muscles, and (4) vocal fold morphological development.

Results. Limited data support the notion that speech tends to produce a contracture of the larynx. The morphological design of the human vocal folds, like that of primates and other mammals, appears to be optimized for vocal communication over distances for which higher fundamental frequency, higher intensity, and fewer unvoiced segments are used.

Conclusion. The positive message is that raising one's voice to call, shout, or sing, or executing pitch glides to stretch the vocal folds, can counteract this trend toward a contracted state.

Key Words: speech–larynx–contracture–singing–muscles.

INTRODUCTION

Mammals, birds, amphibians, and reptiles communicate with sound produced in respiratory airways, using either a larynx or a syrinx (birds) as a sound source. Sonic, infrasonic, or ultrasonic vibration is sustainable by nonlinear interaction between the airstream and a collapsible (soft tissue) segment of the airway wall, producing a fundamental frequency (f_o) and a spectrum of higher frequencies. The vibrating tissue disturbs the airstream so that acoustic waves are propagated in the airways. These waves travel along the slowly moving airstream, with a small portion of the sound being radiated from the mouth, beak, or nostrils into free space to the listener. A much larger portion of the sound is retained in the airway in the form of multiple reflections from irregular boundaries, producing standing waves that form distinct classes of sound. Amplitude and frequency modulations of the f_o and higher partials are added to allow rhythmic and melodic patterns to be produced.

Communication strategies vary across species.[1] A single tone with constant f_o and amplitude, considered the *carrier* of vocal communication, reveals the presence and location of an animal, and possibly its size and some degree of identity. Modulations of this carrier are needed to create a sufficient inventory of sounds for all of the animal's communication needs. In analogy with radio communication theory, the carrier is of a much higher frequency than the modulations, which are in the form of variations in amplitude, fundamental frequency, duration, or frequency spectrum. Ideally, the modulations are not so large that the carrier (voicing) is interrupted. For short-distance communication, however, unvoiced segments have become part of the sound inventory. Prosodic variations, such as melody and accent, do not need to interrupt the carrier and are therefore good candidates for long-range vocal communication. They are used by most species that vocalize. Range of frequency and amplitude, speed of change of frequency and amplitude, and frequency spectrum of the overtones of the carrier frequency, become important criteria for a large inventory of vocal signals that a species can produce.

With the evolution of speech on the order of 100,000 years ago,[2] humans discovered that a larger inventory of modulations could be produced by changing the airway structures rather than simply the sound-source characteristics. Moving the tongue, the lips, the jaw, or the velum allowed the frequency spectrum of the carrier overtones to be modulated amply to produce vowels and consonants. Articulation became the dominant modulation in vocal communication, so much so that whispered speech, or speech with a buzzer held against the neck, is viable today for close-range communication. Source modulations are of secondary importance. The invention of electronic amplification makes source modulations with wide frequency and amplitude ranges even less essential for speech.

The purpose of this paper is to put forth an argument, with existing data never presented in combination, that adaptation of the mammalian larynx for long-range unamplified vocal communication could eventually be reversed with excessive or exclusive use of speech over short distances. Calling, with its many variations (howling, shouting, hooting, roaring, screaming, chanting), is practiced so little that a predisposition to motor control problems in the larynx may exist. Furthermore, infrequent mechanical stretching of laryngeal tissues may diminish the need for a multilayered vocal fold morphology. Fragmentary evidence is given here in the nature of acoustic requirements, vocal fold morphology, vocal fold posturing for speech, and approaches used for voice therapy and training. The methods are in the form of corroborating data sets produced over several decades.

Accepted for publication June 13, 2016.

From the *National Center for Voice and Speech, The University of Utah, Lead Institution, Salt Lake City, Utah; and the †Department of Communication Sciences and Disorders, The University of Iowa, Iowa City, Iowa.

National Center for Voice and Speech, The University of Utah, 136 South Main Street, Suite 320, Salt Lake City, UT 84101-3306. E-mail: ingo.titze@ncvs2.org

Journal of Voice, Vol. 31, No. 2, pp. 135–141

0892-1997

http://dx.doi.org/10.1016/j.jvoice.2016.06.003

METHODS

Acoustic requirements for speech

The ideal sound carrier for speech articulation has a low f_o. A logical argument is that intelligibility of phonemes is improved if source harmonics are closely spaced, given that vocal tract resonances can then be sampled with a greater number of frequencies. With identical formant structure, male speech should by this argument be more intelligible than female speech. This has not been reported, however. In fact, the evidence is somewhat in the opposite direction.[3,4] The answer may lie in the fact that formant frequencies are higher in women than in men, which tends to equalize the sampling issue. At female fundamental frequencies beyond those typically used in speech (ie, singing), it has been shown that intelligibility is indeed reduced.[5–7] There is poorer sampling of the resonances of the vocal tract. For example, because most vowels are identified by the first two resonant frequencies of the supraglottal airway (the exception being rhotic vowels), and because these frequencies center around 500 Hz and 1500 Hz, respectively, a f_o that is not well below 500 Hz provides too few harmonics to sample the resonance peaks effectively. The same can be said for voiced consonants, which also have low resonance frequencies. Thus, there is a strong tendency for humans to keep f_o low during speech. Unvoiced consonants use secondary sound sources with greater spectral density (hisses, clicks, and pops), but these sounds are not carried over distances more than a few meters. They are generally not effective for long-range vocal communication unless amplification is used.

Figure 1 shows the range of sound pressure level (SPL) plotted against range of f_o, known as a voice range profile. The curves labeled loud and soft show the nonspeech range of SPL obtainable 30 cm from the mouth over the entire f_o range of 10 men.[9,10] Superimposed are four speech contours from male classroom teachers who speak daily on the order of 6 hours.[8] Note that 70% of the speech f_o falls below 200 Hz, whereas fundamental frequencies beyond 500 Hz are available from the male larynx with equal or greater SPL variation.

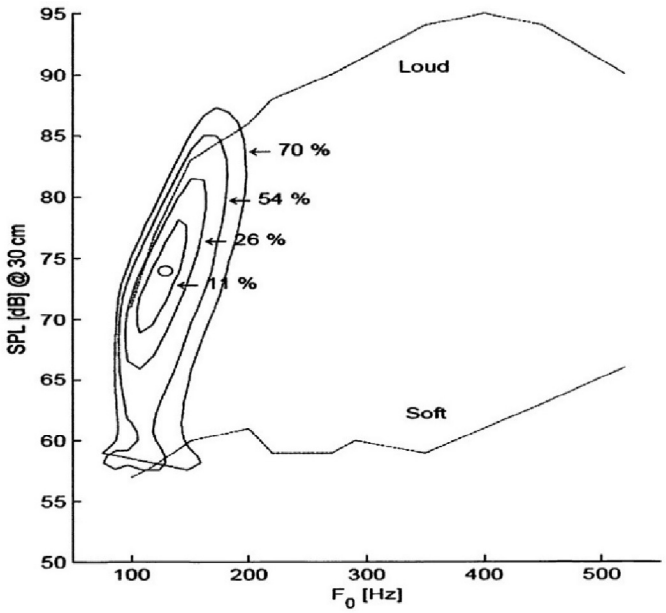

FIGURE 1. Voice range profile with superimposed speech range contours of school teachers (after Hunter and Titze[8]).

Aside from the requirement of low f_o for vowel and voiced consonant intelligibility, strong nonlinear coupling between the source and the filter is avoided when f_o is low. This creates a stability region for source harmonics in speech. It has been shown that low harmonics of the source spectrum (in particular f_o or two f_o) passing through airway resonances can destabilize vocal fold vibration.[11] Avoiding these crossings keeps the source spectrum more constant (less registered in voice science terminology, or less bifurcated in nonlinear dynamics terminology) so that modulation with articulation is not confused with spectral uncertainty at the source. Hence, a linear source-filter theory of speech production that separates the sound source from the resonator has been the hallmark of speech science and technology for more than half a century, notable based on the work of Fant.[12]

A current trend in speech is to use vocal fry, also known as pulse register.[13,14] A subharmonic series (period doubling or tripling) is often produced, which increases the density of source frequencies in the spectrum. This would appear to increase speech intelligibility if the source were to remain stable. Vowels and consonants are extremely well sampled, but overall vocal intensity is low and the voice quality is rough. With amplification or small speaker-listener distances, low intensity does not seem to be an issue. However, prosodic variations (ie, intonation that takes the voice into and out of fry without linguistic intent) appear to be more difficult to produce. Without ample prosodic variation to communicate mood, personality, or other paralinguistic characteristics, speech can degenerate to *vocal texting*.

Before acquisition of the major articulatory components of speech, human infants vocalize at high f_o (around 400–500 Hz). Mothers encourage high f_o vocalization with "mothereses" that modulate the high f_o carrier with lots of prosodic variation.[15] The f_o steadily drops in the first 3 years (to around 300 Hz) as more and more articulation develops.[16] Calling, shouting, laughing, and crying continue toward puberty but are often suppressed thereafter for social reasons. Even singing, which in some cultures is a daily family or school activity, is becoming less habilitated with electronic amplification that can extend the acoustic dimensions of the voice artificially. People who sing appear to retain physiologically younger voices into advanced age. Their voices are louder and their f_o is categorically higher.[17]

Vocal fold morphology

Mammalian vocal folds (or vocal cords) are generally multilayered in their tissue construct (Figure 2). An epithelium (skin) encapsulates a soft-tissue structure known as the lamina propria, which in itself can have one, two, or three compartments: a superficial layer, an intermediate layer, or a deep layer.[19] They consist of a matrix of elastin and collagen fibers with interstitial fluids (proteoglycans and glycoproteins).[20] Lateral to the deep layer, and firmly attached to it, is the thyroarytenoid (TA) muscle.[18]

The design is functional in that the superficial layer (under the skin) needs to be pliable, like a gel, to support surface modes of vibration.[21] These surface modes allow energy to be transferred from the airstream to the tissue. An alternating convergent