Olli Aaltonen & Esa Uusipaikka

Why Speaking Is so Easy?
– Because Talking Is Like Walking with a Mouth

Abstract

One peculiar feature of speech perception is that when we hear sounds, we hear them either as speech or non-speech, and once they are heard as speech we are no longer able to hear them as series of hisses and buzzes, or something in-between, but only as a sequence of vowels and consonants. The human speech perception system works like a sixth sense and, consequently, our impression as listeners is one of immediate, automatic recognition; sounds enter the ear and the result is recognition. The main prerequisite of this uniquely human communication is that the speaker and the listener must have a common understanding that, out of all possible sounds man can produce and hear, only a few have linguistic significance. What counts for a speaker must count for listener, and vice versa, otherwise there is no communication - a fact applicable to all communication, whether linguistic or not. If so, all we have to do to perceive speech is to listen; somehow the meanings just emerge as the sounds go by.

1. Introduction

The fact that speaking is so easy is a consequence of several pre-adaptations for vocal communication being produced by natural selection, some of them arising relatively late in the course of human evolution. As a consequence of this evolutionary process human language became linked to special regions in the brain, the human vocal tract is specially designed for production of syllable-sized complex gestural movements and, finally, our mechanisms of hearing sounds that those gestures produce show precise complementary specializations for decoding the speech produced by other humans. Furthermore, the processes of production and perception became tightly linked via a complex mirror system that matches observation and execution. Thus, the evolved speech system by itself guarantees the important requirement of parity, which is difficult to explain, as the motor

A Man of Measure
Festschrift in Honour of Fred Karlsson, pp. 111–118

acts of the speaker and the auditory percepts of the listener have in common only that neither has anything to do with language.

2. Speech: A special code

There is no doubt that the perception of speech sounds is as effortless and accurate as it is because we are far more experienced in listening to speech sounds than to any other sounds in our environment. What is more in speech comes from the special nature of human speech sounds compared to other acoustically complex signals in nature. For human beings speech sounds are the primary carriers of human language, conveying information about phonemes, which have a special function in language. They correspond to an act of making a sound and serve to mark a change in meaning of an utterance. Being human and being exposed to ambient language are the only necessary conditions for an infant to acquire speech. Thus, learning to speak is an automatic process, in contrast to learning to read or write, which need to be taught in order to be learnt. Imitation plays central role in this process. The meaning of the word plays no part in imitating a word, or a word-like string of sounds. Imitation is a purely formal process where the structure and function of the word are separated from each other.

Usually people who are literate in a language are educated to think that speaking is like writing and that a speaker produces a distinctive acoustic pattern or energy much like a typewriter produces letters, for every distinct vowel and consonant that we perceive. However, this is not the way we speak. Speech is not perceived, produced, or neurally programmed on a segmental basis but utterances are produced and perceived as a whole. In fact, the temporal resolution capacity of the ear would not be good enough at normal speaking rates to segregate different phonemes and to perceive their proper order, if phonemes were consecutive bits of sound (Liberman & Mattingly 1985). Nature circumvents this limit imposed by the auditory system by packing the phonemes in such a way that each segment of sound conveys information about several phonemes. As a consequence, the articulatory gestures characteristic to isolated sounds are never executed in actual speech. Instead they are coalesced into a composite, characteristic of the syllable, and the acoustic properties that specify vowels and consonants change according to the phonetic context.

3. The rival theories: Auditorism and Gesturalism

According to the conventional way of thinking, perception of vowels and consonants has to be auditory, because "we speak to be heard in order to he understood" (Jakobson, Fant & Halle 1963: 13). It is thus tacitly assumed that the phonetic elements are sounds. Therefore, phonetic units cannot be immediate objects of perception; they can only be perceived secondarily, as a result of a cognitive association between a primary auditory representation appropriate to the acoustic stimulus that excites the ear and, on the other hand, some cognitive representation of a linguistic input. This is a basic premise in the auditory theory of speech perception. Auditorists see the perception of speech as a wholly unexceptional example of the workings of an auditory modality that deals with speech just as it does with all the other events we are able to hear. In this respect, the perception of speech is, in principle, no different from the perception of Morse code, or for that matter, the letters of the alphabet.

In contrast, the most extreme, ethologically oriented theory of human speech perception suggests that there is an intimate link between the system responsible for perceiving speech and the system responsible for producing speech (Liberman & Mattingly 1985). The main concept is that we perceive speech by virtue of our tacit knowledge of how speech is produced. Thus, the elements of phonetic structures are gestures, not the sounds those gestures produce. The gestures are ultimately phonetic, having evolved solely for the purpose of phonological communication. Therefore, apprehending phonetic structures has to be managed by a distinct, language-specific system that has its own phonetic domain and its own phonetic mode of processing served by a neurobiology of its own.

Gesturalists claim that nothing would be gained in evolution by a cognitive representation of the communicative elements. Therefore, human and non-human communication alike depends on a specialization at the level of the signal. Thus, the specialized processing system for speech is part of human heritage, operating early in infancy and allowing infants to perceive the sequences of vowel and consonants long before they utter their first words. From this gesturalistic perspective the understanding of the structure and function of the brain is crucial to be compatible with known evolutionary constraints.

4. What purpose the brain is needed for?

Life is based on the existence of variation. Bodies have evolved brains as a buffer against environmental complexity, that is, variation (Allman 1999). Brains help animals, including humans, to co-ordinate behaviour in this continuously varying world in such a way that they can exploit opportunities and avoid hazards. Thus, the brain must have evolved to co-ordinate the actions of the body, and not teleologically to control them. Thus the brain is a *co-ordinator* rather than a *controller*. The brain exists for the body merely to co-ordinate bodily actions in such a way that the purposes of the body can he achieved. Accordingly, the body is the ultimate master instead of being a vehicle that the brain controls for its own purposes. The brain alone does not have any goals.

The brain, especially the cortices, consists of a huge network of neurons and other cells that are chemically and electronically connected. The most important feature of "higher" animals, for example mammals, is that in parts of the brain, especially in neocortex, massive reciprocal connectivity (re-entrant neurons, Edelman & Tononi 2000) is present. Accordingly, brains function as a whole forming ever-changing patterns of activity which may have been triggered by environmental cues via sensory organs, the body's own signals about its state, and needs. The brain is a dynamic pattern generator and the generated patterns co-ordinate the behaviour of the body to achieve its goals, ultimately to survive long enough to reproduce (Kelso 1995). Therefore, the human brain simply cannot be a digital or symbolic computer. From the evolutionary perspective, the brain was not built like a computer with a special design in mind, but natural selection is responsible for its development. Consequently, the behaviour induced by different environmental cues must result from qualitatively different operations in the living brains compared to those in the artificial ones. "People can be calculating, but the brain does not calculate. People can program computers, but brain doesn't program anything, not even the movement of my little finger" (Kelso 1995: 26).

The most important function of animal brains is to co-ordinate movement in the environment in order to get successfully from one place to another or to change the environment. Both behaviours evolved to enhance the survival of an animal. Thus, all animals with brains must have brain systems for navigation in the environment or with the environment. The navigation with places, objects and events is accomplished mainly with automatic and unconscious dynamic patterns produced by various parts of

brains. These patterns are the result of evolution and individual development (Edelman & Tononi 2000, Kelso 1995). Every time when an animal, including humans, perceives a place, an object and/or an event, its brain changes through changes in the synapses of the neurons. This experience is never a passive phenomenon. Perception is always an active process with which any animal by moving, that is, by navigating, actively explores the environment. Therefore, it is very plausible that the categorization of the environment, and ultimately concepts, are realized in brains as dynamic activity patterns connected to movement (schema according to Neisser 1976).

Some animals have more sophisticated bodies to move with than others depending on their natural settings. Having four legs and using them in the same way is in many ways a good thing, but moving in trees changed the style of navigation from four legs to upright posture. Ultimately this produced hands for the primates and, thus, brought at least two new phenomena. The animal could see its body, especially the hands, and could use the hands in different ways compared to those moving in a traditional style (Lorentz 1977: 151). These animals started to use their hands for communication via gestures as a device to strengthen their social bondages in their style of life in groups (presumably as a defence against predators). The social style of life was also based on the ability to imitate, as the result of having mirror neurons.

5. Linking perception and action in speech via mirror neurons

Language has not always existed. According to one scenario the emergence of modern spoken language evolved from earlier pre-phonetic capacity to perform speech sounds or manual gestures. Chimpanzees cannot speak, because to wild chimpanzees voluntarily breath control does not come naturally. On the other hand, chimpanzees have good voluntary control over their manual gestures, although they are not as capable as humans of intricate manual work. A pre-adaptation that was necessary for the emergence of speech was the extension of voluntary control from the hands to the vocal tract (e.g., Calvin 1993). Learning controlled actions by observation entails an ability to imitate. Imitation involves an impressive transformation of sensory impressions into motor commands. A neural basis of imitation has been found in monkeys in the form of mirror neurons, which fire both when an animal is carrying out a certain action,

and when it observes that same action carried out by another animal (Rizzolatti & Arbib 1998).

The mirror system in monkeys is homologue of Broca's area, which is a crucial speech area in humans. This observation provides a neuro-biological missing link for the long-debated hypothesis that primitive forms of communication based on manual gesture preceded speech in the evolution of language. Arbib (2002) suggests that language is carried out by a speech-manual-orofacial gesture complex and a normal child shifts the major information load of language to the speech domain but the mirror system allows for a deaf child the information load of language removed from speech to be taken over by hand and orofacial gestures.

Many of the pre-adaptations for language have been around far longer than modern humans, some of them for hundreds of millions of years while others have existed for considerably shorter time spans. The only latecomer is the human vocal tract and the neural capacity to co-ordinate its various parts (lips, tongue tip/blade, tongue body, tongue root, larynx, and velum) independently of each other (Studdert-Kennedy & Goldstein 2003: 243). At the same time humans developed the ability to co-ordinate breathing in a more flexible way. This latecomer is definitely the key to the phenomenon of speech. It gave modern humans the ability to navigate artificially by moving concretely the parts of vocal tract using the neural system pre-adaptations that had been evolved for moving other parts of the animal bodies. This relocation of the communication from manual gestures to vocal gestures freed the human mind to rise above the immediate needs and, thus, allowed the development of concepts and the way of life we have today. This evolutionary process was not possible to the same extent earlier, because for the early hominids hands had other more important usages which interfered with their usage for communication. The human vocal tract does not have this handicap, because vocal gestures in the mouth do not interfere decisively with other functions of the mouth.

6. Final remarks

According to our scenario, speech is primarily a gestural phenomenon, which evolved from many pre-adaptations for vocal communication in an early man. This development drove changes in physiology and anatomy, allowing vocalisers to control lip muscles independently of tongue muscles, these independently of the soft palate and so on. This particulate principle

supplanted the holistic principle of primate signalling, typically consisting of about thirty calls, each of which is a different holistic sound pattern. In contrast, a typical language is made up of thousands of words, each of which have the same action structure corresponding to the syllable, defined as a vowel surrounded by a consonant. Particulation of the vocal machinery required detailed co-ordination between the active articulators, because none of the articulators work alone; several are active and all are at least passively engaged in every utterance. The complex, now evolved, mirror system allowed for imitation. Imitation gave rise to a new level of processing between a signal and a message, which is lacking from other animals specia1ized in vocal communication. This new level included a phonetic representation and a mechanism for phonetic storage. Finally, all these evolutionary changes in humans made autonomous speech possible, which forms the obligatory basis for thinking and culture and, that is, for being human. In our scenario speech evolved from an ancient animal capacity to navigate. Therefore, speaking is so easy.

References

Allman, John (1999) *Evolving Brains*. New York, NY: Scientific American Library.

Arbib, Michael (2002) The mirror system, imitation, and the evolution of language. In Kerstin Dautenhahn & Chrystopher L. Nehaniv (eds.) *Imitation in Animals and Artifacts*, pp. 229–280. Cambridge, MA: The MIT Press.

Calvin, William H. 1993. The unitary hypothesis: A common neural bases for novel manipulations, language, plan-ahead, and throwing? In Kathleen R. Gibson & Tim Ingold (eds.) *Tools, Language and Cognition in Human Evolution*, pp. 252–278. Cambridge: Cambridge University Press.

Edelman, Gerald M. & Giulio Tononi. (2000) *A Universe of Consciousness: How matter becomes imagination*. New York, NY: Basic Books.

Jakobson, Roman, Gunnar Fant & Morris Halle (1963) *Preliminaries to Speech Analysis.* Cambridge, MA: The MIT Press.

Kelso, J. A. Scott (1995) *Dynamic Patterns: The self-organization of brain and behavior*. Cambridge, MA: The MIT Press.

Liberman, Alvin M. & Ignatius G. Mattingly (1985) The motor theory revised. *Cognition* 21:1–36.

Lorentz, Konrad (1977) *Behind the Mirror: A search for a natural history of human knowledge*. London: Methuen & Co.

Neisser, Ulrich (1976) *Cognition and Reality: Principles and implications of cognitive psychology*. San Francisco, CA: W. H. Freeman & Company.

Rizzolatti, Giacomo & Michael A. Arbib (1998) Language within our grasp. *Trends in Neurosciences* 21.5:188–194.

Studdert-Kennedy, Michael & Louis Goldstein (2003) Launching language. In Morten H. Christiansen & Simon Kirby (eds.) *Language Evolution: States of the art*, pp. 235–254. Studies in the Evolution of Language. New York, NY: Oxford University Press.

Contact information:

Olli Aaltonen
Department of Phonetics
FI-20014 University of Turku
aaltonen(at)utu(dot)fi
http://www.phon.utu.fi

Esa Uusipaikka
Department of Statistics
FI-20014 University of Turku
esa(dot)uusipaikka(at)utu(dot)fi
http://www.soc.utu.fi/tilastotiede