Vocal-auditory feedback and the modality transition problem in language evolution

Sylwester Orzechowski1, Sławomir Wacewicz2, Przemysław Żywiczyński2
1 Institute of Psychology, Maria Curie-Sklodowska University, Lublin, Poland
2 Center for Language Evolution Studies, Nicolaus Copernicus University, Toruń, Poland

Abstract

The two decades of intensive language evolution research have solidified the position of the gestural approach as a major contender in the debate on language origins. Although the gestural theories are intuitively less natural and appealing than the speech-first theories, the arguments for the gestural account are surprisingly compelling and numerous, contributing to the popularity that it now enjoys. Both the advocates and the opponents of the gesture-first position, however, emphasize the gravity of the "modality switch" problem, i.e. how and why language could have transferred from the mostly-visual to the mostly-vocal form that it now has in human linguistic communication. In what follows, we address this problem and suggest its potential solution by appealing to a narrow class of gestures – orofacial gestures, which comprise actions of the muscles of the face and the tongue. With the foundational assumption of a gestural protolanguage, orofacial gestures could be seen as having emerged on the strength of their own communicative potential, piggybacking on preexisting hand-mouth links. Our proposal is that with an increase in the number of gestural signals, including the orofacial ones, the performance of the latter would increasingly rely on the sounds that could accompany their production. This in turn would create pressures on the receiver to map between the vocalizations and their corresponding orofacial gestures as well as pressures on the producer to make the vocalizations maximally distinct, through the operation of vocal-auditory feedback.

Keywords: gesture-first theories, orofacial gestures, modality switch, vocal-auditory feedback, phonological loop

Introduction

The gestural primacy theories (gestural, gesture-first, hand-to-mouth) have come to take center stage in today's academic reflection on the roots of language; put simply, they see the origins of language in a gestural/visual rather than spoken/vocal communicative system. Although the gesture-first position is intuitively less natural and appealing than the speech-first position, the arguments for the gestural account are surprisingly compelling and numerous, contributing to the popularity that it now enjoys. The gestural theories, however, suffer from a near-fatal problem of the so-called "modality switch", i.e. of how and why language could have transferred from the mostly-visual to the mostly-vocal form that it now has in human societies almost universally (e.g. Burling, 2005; Fitch, 2010; Kendon, 2008). In our paper, we offer a potential and partial solution to this problem: we propose a scenario linking the visual modality to the vocal modality and describing a transition from the one to the other. Focusing on one class of gestures - orofacial gestures - we appeal to the notion of total vocal-auditory feedback, and suggest that vocal-auditory feedback may alleviate the modality transition problem.

Gestures

The diversity of both popular and technical meanings associated with the term *gesture* makes this notion an unwieldy one. Because of its phylogenetic dimension, our discussion here is

guided by two perspectives important to language evolution: that of communication between humans and that of primatology.

In human face-to-face communication gestures have a great variety of formal expressions as well as a great variety of functions (cf. Goldin-Meadow, 2003; Kendon, 2004); and again, this richness makes them difficult to taxonomize. Perhaps the most popular classification is that developed by David McNeill (2005, 2012), whereby the various forms of mostly manual movement can be placed on a continuum organised by three variables gradually changing from its left to its right extreme: the co-presence of speech decreases, and both the degree of linguistic properties and of conventionalisation increase (with minor exceptions).

gesticulation - language-slotted gesture - pantomime - emblems/deictics - signs

Thus, gesticulations are the natural and spontaneous movements, mostly of the hand and arm, obligatorily accompanied by speech but lacking any systematic linguistic (e.g. combinatorial) properties or conventionally encoded meanings. On the other hand, the other extreme has the signs of a sign language, which require the absence of speech, are almost fully conventionalised and have linguistic properties equivalent to those that typify the units of a spoken language.

A central concern in the primatological perspective has lain in identifying gestures as those behaviours that are intentionally communicative, as opposed to movements whose nature may be more accidental, unintentional and/or instrumental. For example, Pika (2008) lists the following traits characteristic of gestures proper:

- directed to the receiver,
- mechanically ineffective (i.e. as opposed to mechanically effective instrumental action),
- provoking the desired response,
- performed intentionally.

Since intentionality is difficult to determine, several additional criteria typically hold for asserting it with confidence. They include relative context independence of the behaviour, audience-checking, response-waiting or persistence (e.g. Tomasello, 2008; De Waal & Pollick, 2011).

Here, as elsewhere (e.g. Wacewicz & Żywiczyński, 2008; Orzechowski et al., 2014), we assume a broad definition of gestures, mostly guided by our interest in the question of the communicative modality. Our starting point is the intuitive understanding of gestures as intentional communicative movements of the hand and arm, which we take to be the central - prototypical - examples of that category. However, we extend this to embrace most bodily signals that exist in the visual modality, i.e. are perceived visually, so that “gestures” peripherally include proxemic signals, some intentionally controlled facial displays or even gaze direction.

Of specific interest here are *orofacial gestures*, a subtype of the visible movements of the orofacial area. We have defined orofacial gestures as “movements of the muscles of the front of the head (including the ocular, masticatory, facial and lingual muscles) that are visually accessible to other individuals... [that are] characterized by a high degree of voluntary control, whose production is volitional, flexible, and originates from a communicative intention: they are intended to convey specific information to other individuals” (Wacewicz et al., in press).

The gestural theories of language origins

The so-called gestural theories are a group of theoretical accounts, or “scenarios”, of language origins which postulate that language (more specifically, protolanguage or the earliest forms of language-like communication) began in a gestural communication system, or more broadly, a signaling system of mostly visual nature. Thus, by default, they oppose speech-first theories

(e.g. Dunbar, 1996; Burling, 2005; Mithen, 2005; MacNeilage, 2008). Despite their counterintuitiveness, various forms of gesture-first proposals have a long and rich intellectual history in occidental glossogenetic theorizing.

History

The idea that language originated from gestures, understood widely as communicative body movements, was one of the typical motifs of the naturalistic reflection on language origins in the seventeenth and eighteenth centuries. For example, Vico eloquently argues that the first form of human communication relied on gestures, pictograms, artifacts and religious rituals ([1725] 1948). A better known solution was proposed in the famous thought experiments – by Mandeville (1728) and Condillac (1746) – where a couple of isolated children discover language anew, starting from the natural mode of communication, which in those authors' opinion consists of whole body movements, manual gestures and emotional cries. The scenario whereby the development of language begins with the communication based on the visual component of pantomime/gesture and the vocal component of emotional cries became by far the most popular account of glossogeny in the Enlightenment, which persisted well into the nineteenth century (see e.g. Noiré). At the same time, this motif was solidified by the growing interest in sign languages for the deaf (Amman, Itard, Sicard), traditional sign languages (Jauffret) and the success with which pantomimic communication was used by European travellers with natives from distant lands (Laromiguière). The idea that gestural, or gestural-pantomimic communication, is more universal and hence ancient than its vocal counterpart was inherited by early anthropology and psychology. Tylor untiredly documented traditional signs and emblematic gestures of the cultures he described, which often led him to arguments of how spoken language could have developed from visual communication (1867, 1871, 1881). Finally, Wundt stressed the expressive power of gestures and pantomime, and on this platform reasoned that they represent the primary communicative means both in onto- and phylogeny (1900).

Hewes

The father of modern gestural approach, Gordon W. Hewes had a vast knowledge of the traditional glossogenetic reflection. Combined with his singular reconstructive talent, this brought forth a number of excellent historical outlines (1975, 1976, 1977a, 1996). However, his ambition was to transform glossogenetic theorizing into a truly scientific pursuit, which would follow from empirical data. Although he proposed his own scenario – termed the Gestural Primacy Hypothesis (1973, 1977b), it seems that Hewes's greatest achievement was designating areas of science that could lend best support to gestural theories of language origins. Appealing to the sprouting gesture studies and the Goffmanian micro-analysis, he persuasively argued that natural, face-to-face interaction is multi-modal and relies inasmuch on speech as on gesture (1973). At this juncture, he puts a strong but much needed statement that defining language as a system of purely vocal forms is not supported by interactional facts, but is simply the effect of "the long obsession of linguistics with speech" (1973, p. 11). Another of his insights concerns the discontinuity between language and the primate vocal communication, which he used as indirect evidence supporting the gestural scenario (1973, 1975, 1977a, 1977b). He built this line of argumentation, on the one hand, by documenting failed attempts to train primates in speaking (Furness, 1916; Kellogg & Kellogg, 1933; Hayes & Hayes, 1952) and, on the other, by discussing contemporaneous and promising efforts to use systems of visual signals in teaching language to apes (Gardner & Gardner, 1969, 1971; Premack 1970; Premack & Premack, 1974). Hewes also indicated the potential of neuroscience for the evolution of language in general and for gestural scenarios specifically, calling up evidence from neuropathology and emphasizing the resilience of gestural-pantomimic communication in language-related disorders. Although some of his claims remain controversial, such as the

alleged greater iconicity of sign languages than spoken ones, gestural scenarios found in Hewes an erudite and intelligent supporter, who ensured them a respectable place in the emerging field of language evolution studies.

Contemporary

As mentioned above, the contemporary gesture-first theorists rely on an unexpectedly wide array of arguments - many of them anticipated by Hewes - whose collective force simply cannot be ignored. The most recurrent threads are:

- handedness and lateralisation,
- mirror neurons,
- the iconic potential of gesture, and
- the expressive potential of whole-body mimesis.

In most people, it is the left cerebral hemisphere that is responsible for both the motor control of the dominant upper limb (about 90% of people are right-handed) and for (the bulk of) language processing; a correlation that appears to be systematic (cf. e.g. Knecht et al., 2002). This fact was taken as support for a scenario where lateralisation appeared for motor control of the hand and only later was exapted for speech (Hewes, 1973). Alternatively, language could have lateralised as the largely non-semantic vocalisation – which was already lateralised to the left in non-human primates – was gradually subsuming the more language-like gestural signals that were the original carrier of semantic content (Corballis, 2003). An interesting argument from neuroscience is related to research on mirror neurons. Arbib (e.g. 2005) suggests that the mirror system implements what he calls the 'parity principle': the same form of a signal counts for the same meaning to both the sender and receiver, even though the production and reception processes are different, including their cerebral grounding. Arbib (2005, 2012) develops a sequence of steps that could have led from mirror neurons supporting the understanding of instrumental action (as they do in monkeys) to their gradually extending to embrace non-instrumental, communicative manual gestures.

Another important thread in discussions favoring the gestural scenario is the greater iconic potential of visual than vocal signals. The visual similarity of hand shapes (and hand movements) to the meanings they express could have been a huge cognitive facilitator, creating a kind of natural cognitive link, or 'bridge', between the signal and its referent. This line of thought was developed in most detail by William Stokoe and continuators (Stokoe, 1991; Armstrong et al., 1995; Armstrong & Wilcox 2007). Their account extends the iconicity inherent in gestures to grammar - e.g. the hand functions as a prototypical subject and its movement - as a prototypical verb, making one gesture holistically represent a complete sentence. Related to iconicity is the expressive potential of whole-body mimesis (Donald, 1991; Zlatev 2008, 2014), i.e. conveying messages with the movement of the entire body, which could be used to represent objects but also reenact complex events. Mimetic communication can be taken to be 'simpler' and more primitive than language as it relies on more 'direct', iconic meanings without the need to refer to conventional, socially negotiated symbols. Characteristically, on many of the accounts mentioned above, the original protolanguage would have been in whole body pantomime rather than strictly manual gesture: not only Donald and Zlatev, but also Arbib, and Tomasello (2008) envisage a more or less pantomimic stage in language origins (for criticism see McNeill, 2016).

The modality transition problem

The main problem of the gesture-first accounts of language origins can be termed the *modality switch* or *modality transition problem*: If language arose as a (predominantly) gestural/visual system, why would it have changed to its present (predominantly) spoken/vocal form, which is

backed up by the extensive anatomical and neuroanatomical human adaptations to speech production?

This difficulty did not go unnoticed by the supporters (e.g. Hewes, 1973) as well as critics (e.g. MacNeilage 2008). As stated by Fitch (2010, p. 434):

> [A] significant disadvantage of gestural models is their difficulty in explaining the virtually complete transition to vocal, spoken language in modern Homo sapiens... Whatever their virtues, models of gestural protolanguage are incomplete without a detailed and compelling model of the transition to spoken language, as most gestural proponents have recognized…

The same is observed by Kendon (2008, p. 12):

> Yet, as has often been pointed out, this seemingly attractive hypothesis faces, as MacNeilage (1998, p. 232) has put it, an insuperable problem. Languages are overwhelmingly spoken. Furthermore, humans appear to be specialised anatomically to be speaking animals….

Solutions

Language origins literature is replete with observations that point to one or another advantage of a spoken relative to gestural communication system. Although superficially such observations may seem to speak against the idea of gestural primacy, considered specifically in the context of the *modality transition problem*, they could justify selection pressures for such a transition. Unfortunately, most such points have a loose and anecdotal character and are difficult to develop into more complete and compelling conceptual arguments, let alone translate into testable predictions. We list them below as interesting observations, but we wish to note that as arguments for explaining the putative gesture to speech transition, they are simply insufficient.

- Speech is energetically more economic than manual gesture (e.g. Knight, 2000);
- speech makes it possible to communicate in the dark, and sound is more efficient at attracting attention (those observations date back at least to Rousseau [1775]);
- speech frees the upper limb for tool use, toolmaking, or other manual instrumental activity (e.g. Carstairs-McCarthy, 1996);
- speech makes it possible to comment on, and thus explicitly teach, manual action (Armstrong and Wilcox, 2007);
- the acquisition of spoken language begins in the foetal life, which gives that modality a developmental advantage (Hewes, 1996);
- vocal communication makes it possible for the mother to monitor the position of the infant, which unlike in the other apes does not remain in obligatory physical contact with the mother (Falk, 2009);
- vocal communication makes it possible to address many individuals at once (e.g. Tomasello, 2008)

Fitch (2010) discusses many of the above points, and rather convincingly questions their strength and validity in explaining the 'modality transition'. For example, gestures, while invisible in the dark, are visible in the firelight, and they can also be perceived haptically, a mode of communication practised today by deaf signers. The visual channel is also more efficient in the noise or over very long distances; and while gesture occupies the hands, speech occupies the oral cavity, which would have been used by our palaeolithic ancestors much more extensively than today not only for chewing (necessary for undomesticated and unprocessed food) but also performing instrumental mechanical action. Fitch further notes that the energetic efficiency argument fails, too, because in spontaneous conversation the vocal messages almost invariably

involve co-present gesticulation, which makes this way of communicating at least as energetically costly as pure gesture.

The other arguments mentioned above, but not considered by Fitch, are likewise unconvincing. In the teaching of complex manipulation verbal instruction may be of some help (Morgan et al., 2015), but is not as effective as demonstration or physical guiding of the hands of the disciple. Hewes' point about speech acquisition *in utero* appears to be of rather marginal importance, especially in the light of the more recent developmental data which show that sign languages tend to be acquired on a parallel, if not slightly faster time course (Petitto, 1994). Falk's speculation, interesting as it is, does not need to involve any language-like system (combinatorial, semantic, etc.) but only generic emission of noise. Tomasello's observation is accurate, however this point may be countered by the greater secrecy of gestural communication, allowing the producer to strategically choose the addressees of the message and exposing them to a lower risk of detection by predators or competitors (Wacewicz & Zywiczynski, 2008).

Orofacial solutions

The idea that orofacial gestures form a platform for accomplishing the transition has an interesting history. Its influential modern supporters (Woll, 2014; Meguerditchian et al., 2014) go back to Darwin's cursory observations of how the tongue movements accompany bodily routines, particularly fine hand actions (1872). The first proponent of what could be called the "orofacial hypothesis", i.e. the hypothesis that orofacial gestures formed a bridge facilitating the transition between gestural and vocal communication, was the co-author of the natural selection theory, Alfred Wallace (1881, 1895) (although Woll credits the phonologist Sweet [1888] with the first explicit mention of the hypothesis; see Woll, 2014). Probably, the staunchest but also most controversial advocate of the orofacial hypothesis was R. A. S. Paget. His mouth gesture theory opens with Darwin's observation, enriched by Paget with numerous new examples, that the mouth and other articulators often echo hand movements. This led him to the radical statement that spoken language arose in the process of the mouth, tongue and lips involuntarily imitating body movements (Hawhee, 2006). What is important in this context is that, unlike Wallace, Paget was not interested in the visual signaling potential of the orofacial area; what mattered to him were the acoustic consequences of the mouth and tongue movements, or "lip-reading by ear" as he called it: "The significant elements in human speech are the postures and gestures [of the organs of articulation], rather than the sounds. The sounds only serve to indicate the postures and gestures which produced them. We lip-read by ear" (1930, p. 174).

In contemporary language evolution, the orofacial hypothesis has returned, supported by a variety of arguments coming from anthropology (Hewes, e.g. 1996), linguistics (Studdert-Kennedy, 2002), and primatology (e.g. Meguerditchian et al., 2011). The most comprehensive account comes from Corballis (2002, 2003, 2012), who highlights the fact that monkeys and non-human apes possess voluntary control over manual as well orofacial gestural actions afforded by neocortical connections, but lack such control over their vocalizations. This fact is used by Corballis to build the argument – similar to that of Paget – about orofacial gestures as a platform via which (proto)language could have transitioned from the manual-visual to the vocal-auditory modality. Again, in consonance with Paget, Corballis presses the point that speech is essentially a system of movements of the speech organs, and may thus be considered a system of "gestures". This is also reminiscent of the conception developed by Armstrong, Stokoe and Wilcox (1994, 1995; also Armstrong & Wilcox, 2007), who view both gestures and speech as "planned sequences of musculo-skeletal actions" (for criticism see Kendon 2008). The continuity of manual and speech gestures is underscored by the addition of vocalization, which makes many of the orofacial movements accessible to the receiver in the vocal-auditory modality. On this scenario, the relatively flexible orofacial area plays a major role in the gradual evolutionary

extension of flexible voluntary control to the more internal parts of the vocal tract (Wacewicz et al. in press).

The orofacial hypothesis forms an integral part of the gestural-pantomimic scenario developed by Michael Arbib (2002, 2005, 2006, 2012). The ramifications of the Mirror System Hypothesis (MSH) are that vocalization could have been recruited by the original communicative system constituted of both manual and orofacial gestures. Arbib's argument is based on the discovery of the mirror neuron system; he speculates that a corresponding structure in the hominin brain could have provided the neural groundwork for volitionally controlled vocalizations. Arbib's scenario gains support from primatologists such as Leavens, Taglialatela and Hopkins (2014) or Meguerditchian and Vauclair (2014), who, appealing to comparative data, argue that "the oro-facial system might constitute a relevant mediator between the gestural communicatory system and speech in the evolution of language" (Meguerditchian & Vauclair 2014, p. 148).

Vocal-auditory feedback and the modality transition problem

If we accept that orofacial gestures could have acted to facilitate the modality transition discussed above, it is interesting to pay closer attention to the role of *vocal-auditory feedback* in this process. Specifically, we focus on the *sender* rather than receiver of messages, and on the gains in the control of production following from being able to hear one's own voice. While our argument is limited in presupposing some version of the orofacial scenario, it may help explain the problem of modality transition in this 'local' context.

Our logic here partly coincides with the line of thinking outlined by Corballis (2002, p. 185):

> [...] language evolved as a system of gestures based on movements of the hands, arms, and face, including movements of the mouth, lips, and tongue. It would not have been a big step to add voicing to the gestural repertoire, at first as mere grunts, but later articulated so that invisible gestures of the oral cavity could be rendered accessible, but to the ear rather than the eye.

It seems that in this scenario Corballis highlights the benefits for the receiver, who could now rely on sound as well as vision in the process of decoding the message. In particular, as the complexity of the communication system increases and the inventory of discrete signals grows larger, the signal space becomes crowded, and telling apart the subtle differences between the items becomes progressively more difficult. Being able to rely on redundantly structured, bimodal (vision and sound) rather than non-redundant unimodal (sound only) signal makes the messages more robust and easier to comprehend.

Here, in contrast, we focus on the process of production rather than comprehension. Working from the assumptions presented above, we wish to suggest that complementing orofacial gesture with sound would have benefited not only the receiver of the messages, but first and foremost the producer: the addition of auditory access - *vocal-auditory feedback* - would have enabled a more precise control over the execution of orofacial movements.

<u>The feedback mechanisms</u>

Hockett (1960) proposed *total feedback* as one of the design features of language. By this he meant that the senders of the signal (which he assumed to be speech by default) receive full auditory information on the signals they produce, i.e. hear themselves speak. In fact, when considering speech production in neural terms, the problem of feedback is much more complex. The precise neural control required for it depends on a number of feedback circuits, which expedite the coordination of phonatory and articulatory muscles. The most important of these is the auditory feedback circuit responsible for making online corrections of speech sounds with

respect to their representations as neural sound templates. The activity of auditory feedback is supported by the orosensory feedback loop (e.g. Guenther & Perkell, 2004), which consists of kinaesthetic feedback, based on proprioceptive information related to the activity of the speech muscles (e.g. Markides, 1983) and tactile feedback transmitting sensations mainly from the tongue and lips (e.g. Markides, 1983). Hence, speech production is serviced by multimodal feedback - auditory as well as proprioceptive and tactile. It is also interesting to observe that enhancing the process of speech training with *visual feedback* helps articulation: Katz and Mehta (2015) found that viewing the movements of one's own speech organs using a 3D visualization led to improved articulatory accuracy of non-native speech sounds in adult learners.

Auditory feedback

Although in the present context we use the terms auditory feedback and vocal-auditory feedback somewhat interchangeably, in principle *auditory feedback* is more general, being related to any type of own sound production (including e.g. sounds of locomotion or playing an instrument). *Vocal-auditory feedback* is a subcategory: the ability of the speaker to monitor their *vocal* production. Vocal-auditory feedback is crucial not only for the maintenance of a stable internalised speech model (Brainard & Doupe, 2000; Johnes & Munhall, 2003) but also for its acquisition during the vocal development, when it allows learners to flexibly shape motor programmes (or templates) for producing speech sounds (e.g. Borden, 1979; Oller & Eilers, 1988; Osberger & McGarr, 1982; Smith, 1975). Interestingly, vocal-auditory feedback performs a comparable role in the acquisition of song-vocalisations by song-birds (e.g. Brainard & Doupe, 2000), and is of paramount importance in music, where having auditory access to one's own musical production is indispensable for successful performance, both in singing and playing an instrument. On clinical grounds, postlingually deafened individuals manifests abnormalities in the control of pitch, loudness, the rate of speech, increased variability in consonant and vowel production (Binnie et al., 1982; Cowie & Douglas-Cowie, 1992; Lane & Webster, 1991; Waldstein, 1990). As already indicated, the primary function of vocal-auditory feedback is related to making online corrections of speech (Johnes & Munhall, 2003). This was documented in studies using the Delayed Auditory Feedback Paradigm (DAF) (Lee, 1950a, 1950b, 1951), which concluded that disruption of vocal-auditory feedback results in substantial articulatory distortions (Yates, 1973). Similarly, masking vocal-auditory feedback with noise affects pitch (e.g. Rivers & Rastatter, 1985; Ternström et al., 1988), while manipulating the spectra of feedback by raising or lowering F_0 (Garber & Moller, 1979) leads to the compensatory activity of shifting pitch in the direction opposite to the perturbation (Burnett et al., 1998; Donath et al., 2002; Elman, 1981; Jones & Munhall, 2000, 2002; Kawahara, 1999a, 1999b).

The scenario

Our idea for the modality transition is rooted in Corballis' way of thinking about the primary mode of communication - a gestural protolanguage - and the preexisting hand-mouth links (2002). On this view, it is easy to see how orofacial gestures would have emerged on the strength of their own communicative potential, piggybacking on the hand-mouth links. We propose that with an increase in the number of gestural signals, including the orofacial ones, the performance of the latter would rely to a greater and greater degree on accompanying vocalizations. This in turn would create pressures on the receiver to map between the vocalizations and their corresponding orofacial gestures as well as pressures on the producer to make the vocalizations maximally distinct, through the operation of vocal-auditory feedback. On the neural level, this trend towards reliance on vocalization could be explained by the existence of mirror neurons for orofacial movements (Buccino, 2004).

As noted above, speech production depends not just on vocal-auditory but multi-modal feedback. Similarly, in gesturing and signing, the sender of the signal receives multi-modal

feedback, that is proprioceptive as well as visual information on one's hand and arm movements, with some marginal tactile information. The situation is different in the case of orofacial gestures, where the visual feedback is not possible and the producer is guided solely by the proprioceptive/tactile information. A large inventory of communicative signals leads to the crowding of the signal space; as a result differentiation between the signals in their production becomes increasingly error-prone, i.e. it becomes more and more difficult to reliably produce the fine distinctions between the growing number of gestural items. The production of sound together with a specific item (orofacial configuration) would have added vocal-auditory feedback to existing proprioceptive feedback, thus making the configurations easier to control. As a next step, the development of stable pairings between a particular orofacial configuration and a particular vocal signal would have made it possible for the vocal to take over as the 'signifier'. That is, the vocal signal alone could be used to reliably identify the meaning previously expressed in the visual modality by the orofacial gesture (cf. Orzechowski et al., 2014).

Finally, we may note that the function of vocal-auditory feedback extends beyond the discriminatory role described in the first step of our scenario, but may have helped in the subsequent development of spoken communication by being an essential element of the phonological loop. The phonological loop is a subcomponent of working memory that specializes in the retention of verbal information over short periods of time, and comprises phonological store (retaining information in the phonological form) and rehearsal process (maintaining decaying representations in the phonological form) (Baddeley et al., 1998). Importantly, the phonological loop functions to assist the learning of new lexical labels, i.e. to help generate long-term representations of novel phonological material (Baddeley et al., 1998): "… the primary purpose for which the phonological loop evolved is to store unfamiliar sound patterns while more permanent memory records are being constructed" (Baddeley et al., 1998). *Via* the phonological loop, vocal-auditory feedback would have worked to solidify phonological patterns and thus stabilize the "vocabulary".

Conclusion

The communicative potential of orofacial gestures has been acknowledged by language evolution researchers (e.g. Studdert-Kennedy, 2005; MacNeilage, 2008; see also Paget, 1930), who frequently comment on their relevance to the "modality transition" problem that plagues the gestural theories of language origins. To recapitulate our argument, we take as our starting point a gestural scenario on which emerging language-like communication involves orofacial gestures, and we complement such a scenario with the inclusion of vocal-auditory feedback, which aids signal production; this might have been the driving force behind the increasing role of vocalization in linguistic messages. Our proposal is distinct from those of other authors by the fact that we see *vocal-auditory feedback* as the main mechanism: supplementing orofacial gestures with sound is not selected primarily for its benefits to the *receivers*, but rather to the signal *producers*. Those hypotheses need not be perceived as exclusionary; they are best viewed as potentially complementary but distinct solutions. Although admittedly local and limited in scope to the domain of orofacial gestures, our proposal may alleviate the 'modality transition' problem for the gestural models of language emergence.

References

Arbib, M. (2002). The mirror system, imitation, and the evolution of language. In K. Dautenhahn & Ch. Nehaniv (Eds.), *Imitation in animals and artifacts. Complex adaptive systems* (pp. 229-280). Cambridge, MA: MIT Press.

Arbib, M. (2005). From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences*, *28*, 105–167.

Arbib, M. (2006). The Mirror System Hypothesis on the linkage of action and Languages. In M. Arbib (Ed.), *Action to Language via Mirror Neuron System* (pp. 3-47). Cambridge: Cambridge University Press.
Arbib, M. (2012). *How the brain got language*. Oxford: Oxford University Press.
Armstrong, D., Stokoe, W. C. & Wilcox, S. E. (1994). Signs of the origin of syntax. *Current Anthropology*, *35(4)*, 349-368.
Armstrong, D., Stokoe, W. C. & Wilcox, S. E. (1995). *Gesture and the Nature of Language.* Cambridge: Cambridge University Press.
Armstrong, D., & Wilcox, S. E. (2007). *The gestural origin of language*. Oxford: Oxford University Press.
Baddeley, A. D., Gathercole, S. E. & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, 105, 158-173.
Binnie, C. A., Daniloff, R. G, Buckingham, H. W. jr (1982). Phonetic disintegration in a five-year-old following sudden hearing loss. *Journal of Speech and Hearing Disorders*. 47(2), 181-189.
Borden, G. J. (1979, May). An interpretation of research of feedback interruption in speech. *Brain and Language*, 7, 307-319.
Brainard, M. S., & Doupe, A. J. (2000). Auditory feedback in learning and maintenance of vocal behaviour. *Nature Reviews Neuroscience*, 1, 31–40.
Buccino, G., Binkofski, F. & Riggio, L. (2004). The mirror neuron system and action recognition. *Brain and Language*, 89, 370-376.
Burling, Robbins (2005). *The talking ape: How language evolved*. Oxford: Oxford University Press.
Burnett, T. A., Freedland, M. B., Larson, C. R. & Hain, T. C. (1998). Voice f0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America*, 103, 3153-3161.
Carstairs-McCarthy, A. 1996. Review of D. Armstrong, W. C. Stokoe and S. Wilcox: “Gesture and the nature of language”. *Lingua*, 99: 135-8.
Condillac, E. B. 1746. Essai sur l’origine des connaissances humaines. Paris.
Corballis, M. C. (2002). *From hand to mouth: The origins of language*. Princeton, NJ: Princeton University Press.
Corballis, M. C. (2003). From mouth to hand: gesture, speech, and the evolution of right-handedness. *Behavioral and Brain Sciences*, 26, 199-208.
Corballis, M. C. (2012). How language evolved from manual gestures. *Gesture*, 12, 200-226.
Cowie, R. & Douglas-Cowie, E. (1992). *Postlingually acquired deafness: speech deterioration and the wider consequences*. Berlin: Mouton de Gruyter.
Darwin, Ch. (1872). *The expression of emotions in man and animals*. London: Murray.
de Waal, F. B. M. & Pollick, A. S. (2011). Gesture as the most flexible modality of primate communication. In K. R. Gibson & Maggie Tallerman (Eds.), *The Oxford Handbook of Language Evolution* (pp. 82-89). Oxford: Oxford University Press.
Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Cambridge, Mass.: Harvard University Press.
Donath, T. M., Natke, U. & Kalveram, K. T. (2002). Effects of frequency-shifted auditory feedback on voice f0 contours in syllables. *The Journal of the Acoustical Society of America*, 111, 357–366.
Dunbar, R. (1996). *Grooming, gossip and the evolution of language*. London: Faber & Faber.
Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *The Journal of the Acoustical Society of America*, 70(1), 45-50.
Falk, D. (2009). *Finding our tongues: Mothers, infants and the origins of language*. New York: Basic Books.
Fitch, W. T. (2010). *The Evolution of Language*. Cambridge: Cambridge University Press.

Furness, W. H. (1916). Observations on the mentality of chimpanzees and orangutans. *Proceedings of the American Philosophical Society,* 281-290.
Garber, S. R. & Moller, K. (1979). The Effects of Feedback Filtering on Nasalization in Normal and Hypernasal Speakers. *Journal of Speech, Language, and Hearing Research*, 22, 321-333.
Gardner, R. A. & Gardner, B. T. (1969). Teaching Sign Language to a Chimpanzee. *Science,* 165(3894), 664–672.
Gardner, B. T. & Gardner, R. A. 1971 Two-way communication with an infant chimpanzee. In. In A. Schrier and F. Stollnitz (Eds.), *Behaviour of nonhuman primates* (vol. 4, pp. 117-184). New York: Academic Press.
Goldin-Meadow, S. & Singer, M. A. (2003). From children's hands to adults' ears: Gesture's role in teaching and learning. *Developmental Psychology,* 39(3), 509-520
Guenther, F. & Perkell, J. (2004). A Neural Model of Speech Production and its Application to Studies of the Role of Auditory Feedback in Speech, In B. Maassen, R. Kent, H.F.M. Peters, P. Van Lieshout & W. Hulstijn (Eds.), *Speech Motor Control in Normal and Disordered Speech* (pp. 29-50). Oxford University Press,.
Hawhee, D. (2006). Language as Sensuous Action: Sir Richard Paget, Kenneth Burke, and Gesture-Speech Theory. *Quarterly Journal of Speech*, 92, 331-354.
Hayes, K. & Hayes, C. (1952). Imitation in a home-raised chimpanzee. *Journal of Comparative and Physiological Psychology*, 45, 450-459.
Hewes, G. W. (1973). Primate communication and the gestural origin of language. *Current Anthropology*, 14, 5-24.
Hewes, G. W. (1975). *Language origins: a bibliography*. The Hague: Mouton.
Hewes, G. W. (1976). The current status of the gestural theory of language origin. *Annals of the New York Academy of Sciences*, 280, 482-504.
Hewes, G. W. (1977a). A Model for Language Evolution. *Sign Language Studies*, 15, 97-168.
Hewes, G. W. (1996). A history of the study of language origins and the gestural primacy hypothesis. In A. Lock & Ch. Peters (Eds.), *Handbook of human symbolic evolution* (pp. 263-269). Oxford: Oxford University Press.
Hockett, Ch. F. (1960). The origin of speech. *Scientific American*, 203, 88–111.
Jones, J. A. & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. Journal of the Acoustical Society of America, 108, 1246-1251.
Jones, J. A. & Munhall, K. G. (2002). The role of auditory feedback during phonation: Studies of Mandarin tone production. Journal of Phonetics, 30, 303-320.
Jones, J. A. & Munhall, K. G. (2003). Learning to produce speech with an altered vocal tract: The role of auditory feedback. Journal of the Acoustical Society of America, 113, 532-543.
Katz, W. F., & Mehta, S. (2015). Visual Feedback of Tongue Movement for Novel Speech Sound Learning. *Frontiers in Human Neuroscience*, 9, 612. http://doi.org/10.3389/fnhum.2015.00612
Kawahara, H., Katayose, H., De Cheveigné, A., & Patterson, R. D. (1999a). Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. *EuroSpeech*, 99(6), 2781-2784.
Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999b). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*. 27, 187–207.
Kellogg, W. N. & Kellogg, L. A. (1933). *The ape and the child*. New York: Hafner.
Kendon, A. (2004). *Gesture. Visible Action as Utterance*. Cambridge: Cambridge University Press.
Kendon, A. (2008). Signs for Language Origins? *The Public Journal of Semiotics*, 2, 2-29.

Knight, Ch. (2000). Play as precursors of phonology and syntax. In Ch. Knight, M. Studdert-Kennedy & J. Hurford (eds.). *The Evolutionary Emergence of Language* (99-119). Cambridge: Cambridge University Press.
Knecht, S., Flöel, A., Dräger, B., Breitenstein, C., Sommer, J., Henningsen, H., Ringelstein, E.B., Pascual-Leone, A. (2002). Degree of language lateralization determines susceptibility to unilateral brain lesions. *Nature Neuroscience*. 5(7), 695-699.
Lane, H. & Webster, J. (1991) Speech deterioration in postlingually deafened adults, *Journal of the Acoustical Society of America*, 89, 859-866.
Leavens, D., Taglialatela, J. & Hopkins, W. (2014). From grasping to grooming to gossip: innovative use of chimpanzee signals in novel environments supports both vocal and gestural theories of language origins. In M. Pina & N. Gontier (Eds.), *The evolution of social communication in primates: A multidisciplinary approach* (pp. 179-194). New York: Springer.
Lee, B. S. (1950a). Effects of delayed speech feedback. *Journal of the Acoustical Society of America*, 22 (6), 824-826.
Lee, B. S. (1950b). Some effects of side-tone delay. *Journal of the Acoustical Society of America,* 22 (5), 639-640.
Lee, B. S. (1951). Artificial stutter*. Journal of Speech and Hearing Disorders*, *16, 53-55.*
MacNeilage, P. F. (1998) The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499-511.
MacNeilage, P. F. (2008). *The Origin of Speech*. Oxford: Oxford University Press.
Mandeville, de, B. (1728). *The Fable of the Bees.* Part II. London: J. Roberts.
Markides, A. 1983. The Speech of Hearing Impaired Children. Manchester: Manchester University Press.
McNeill, D. (2005). *Gesture and Thought*. Chicago: University of Chicago Press.
McNeill, D. (2012). *How Language Began. Gesture and Speech in Human Evolution*. Cambridge: Cambridge University Press.
McNeill, D. (2016). *Why We Gesture: The Surprising Role of Hand Movements in Communication.* Cambridge: Cambridge University Press.
Meguerditchian, A., Cochet, H., & Vauclair, J. (2011). From gesture to language: ontogenetic and phylogenetic perspectives on gestural communication and its cerebral lateralization. In A. Vilain, J.-L. Schwartz, Ch. Abry & J. Vauclair (Eds.), *Primate communication and human language: vocalization, gestures, imitation and deixis in humans and non-humans* (pp. 89-118). Amsterdam: John Benjamins.
Meguerditchian, A., Plouvier, M., Pruetz, J. & Hopkins, W. (2014). From hand to mouth: fine precision grip during mutual grooming elicited wide lip movements in wild fongoli chimpanzees. In E. Cartmill, S. Roberts, H. Lyn & H. Cornish (Eds.), *The Evolution of Language. Proceedings of the 10th International Conference (EVOLANG 10)* (pp. 487-488). Singapore: World Scientific.
Meguerditchian, A. & Vauclair, J. (2014). Communicative Signaling, Lateralization and Brain Substrate in Nonhuman Primates: Toward a Gestural or a Multimodal Origin of Language? *Humana Mente Journal of Philosophical Studies*, 27, 135-160.
Mithen, S. (2005). *The singing Neanderthals: the origins of music, language, mind and body*. London: Weidenfeld & Nicholson.
Morgan, T. J. H., Uomini, N. T., Rendell, L. E., Chouinard-Thuly, L., Street, S. E., Lewis, H. M., … Laland, K. N. (2015). Experimental evidence for the co-evolution of hominin tool-making teaching and language. *Nature Communications*, 6, 6029, DOI: 10.1038/ncomms7029.
Oller, D. K., Eilers, R. E. (1988). The role of audition in infant babbling. *Child Development*. 59(2), 441-449.
Orzechowski, S., Wacewicz, S., & Żywiczyński, P. (2014). Orofacial gestures in language evolution. The auditory feedback hypothesis. In E. Cartmill, S. Roberts, H. Lyn & H. Cornish

(Eds.), *The Evolution of Language. Proceedings of the 10th International Conference (EVOLANG 10)* (pp. 221-227). Singapore: World Scientific.
Osberger, M. J., & McGarr, N. S. (1982). Speech production characteristics of the hearing impaired. In N. Lass (ed.) *Speech and language: Advances in basic research and practice*, 257-316). New York: Academic Press.
Paget, Richard A. S. (1930). *Human Speech: Some Observations, Experiments, and Conclusions as to the Nature, Origin, Purpose and Possible Improvement of Human Speech*. London: Kegan Paul & Trench Trübner.
Petitto, L. A. (1994). Are signed languages "real" languages? Evidence from American Sign Language and Langue des Signes Québécoise. Reprinted from: *Signpost* (*International Quarterly of the Sign Linguistics Association*) 7(3), 1-10.
Pika, S. (2008). What is the nature of the gestural communication of great apes? In: J. Zlatev, T. Racine, C. Sinha i E. Itkonen (eds.), *The Shared Mind: Perspectives on intersubjectivity* (pp. 165-186). Amsterdam: John Benjamins Publishing.
Premack, D. 1970. A functional analysis of language. *Journal of the Experimental Analysis of Behavior*, 14, 107-125
Premack, D. & Premack, A. J. (1974). Teaching visual language to apes and language-deficient persons. In R. L. Schiefelbusch & L. L. Loyd (Eds.), *Language perspectives: acquisition, retardation and intervention* (pp. 347-375). Baltimore: University Park Press.
Rivers, C. & Rastatter, M. P. (1985). The effects of multitalker and masker noise on fundamental frequency variability during spontaneous speech for children and adults. *The Journal of Auditory Research*, 25, 37-45.
Rousseau, J.-J. (1775). *Discours sur l'origine et les fondements de l'inégalité parmi les hommes*. Amsterdam: Marc Michel Rey.
Smith, C. R. (1975). Residual Hearing and Speech Production in Deaf Children. *Journal of Speech, Language, and Hearing Research*. 18, 795-811.
Stokoe, W. C. (1991). Semantic phonology. *Sign Language Studies*, 71, 107-114.
Studdert-Kennedy, M. (2002). Mirror neurons, vocal imitation and the evolution of particulate principle. In M. Stamenov & V. Gallese (Eds.), *Mirror neurons and the evolution of brain and language* (pp. 207-227). Amsterdam: John Benjamins,.
Studdert-Kennedy, M. (2005) How Did Language go Discrete? In M. Tallerman, (ed.) *Language Origins: Perspectives on Evolution* (pp. 48-67). Oxford: Oxford University Press.
Sweet, Henry (1888*). A history of English sounds from the earliest period, with full word-lists*. Oxford: Clarendon (retrieved from https://archive.org/details/ahistoryenglish05sweegoog).
Ternström, S., Sundberg, J. & Colldén, A. (1988). Articulatory F0 perturbations and auditory feedback. *Journal of Speech and Hearing Research*, 31(2), 187-192.
Tomasello, Michael. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.
Tylor, E. (1867). On traces of the early mental condition of man. *Notes on the Proceedings at the Meetings of the Royal Institution of Great Britain*, 5, 83-93.
Tylor, E. (1871). *Primitive Culture*. Vol. I & II. London: John Murray.
Tylor, E. (1881). *Anthropology an introduction to the study of man and civilization*. London: Macmillan.
Vico, G. (1725/1948). [Scienza Nuova di Giambattista Vico] The New Science of Giambattista Vico. Translation from the third edition (1744) Thomas G. Bergin, M. H. Fish. Ithaca, NY: Cornell University Press.
Wacewicz, S. & Żywiczyński, P. (2008). Broadcast Transmission, Signal Secrecy and Gestural Primacy Hypothesis. In A. D. M. Smith & K. Smith and R. Ferrer-i-Cancho (eds.), *Proceedings of the 7th International Conference on the Evolution of Language* (pp. 354-361). World Scientific.
Wacewicz, S., Żywiczyński, P. & Orzechowski, S. (in press). Visible movements of the orofacial area: evidence for gestural or multimodal theories of language evolution? *Gesture*.

Wallace, Alfred .R. (1881). Review of Tylor's Anthropology. *Nature*, 24, 242-245.
Wallace, Alfred R. (1895). The Expressiveness of speech, or mouth gesture as a factor in the evolution of speech. *Fortnightly Review*, 58, 528-543.
Waldstein, R. S. (1990). Effects of postlingual deafness on speech production: implications for the role of auditory feedback. *The Journal of the Acoustical Society of America*, 88(5), 2099-2144.
Woll, B. (2014). Moving from hand to mouth: echo phonology and the origins of language. *Frontiers in Psychology*, 5, article 662.
Wundt, W. (1900). Völkerpsychologie. Vol. I-II. *Die Sprache*. Leipzig: Engelmann.
Yates, A. J. (1963). Delayed auditory feedback. *Psychological Bulletin* 60(3), 213-232.
Zlatev, J. (2008). The co-evolution if intersubjectivity and bodily mimesis. In J. Zlatev, T. Racine, C. Sinha & E. Itkonen (Eds.), *The Shared Mind: Perspectives on intersubjectivity* (215-244). Amsterdam: John Benjamins.
Zlatev, Jordan (2014). Human Uniqueness, Bodily Mimesis and the Evolution of Language. *Humana Mente. Journal of Philosophical Studies*, 27, 197-219.