The Dissociation of Consonants and Vowels from CV Frames in the Phylogeny of Language

Jarosław Weckwerth
Adam Mickiewicz University, Poznań, Poland
E-mail: wjarek@ifa.amu.edu.pl

ABSTRACT

This paper investigates the origins of the compositional principle in phonology. MacNeilage's [12] "Frame/Content" theory is used as a point of departure. The paper presents a scenario in which stabilisation of production resulting from (1) auditory and tactile feedback gained during exploration of the articulatory space and (2) attempts at imitation of ambient sound structures played a crucial role in the process of digitisation of phonology. It is argued that, since acoustic correlates of individual articulatory gestures are always bundled in acoustic feedback, gesture bundles possibly corresponding to modern phonetic segments may have been used in the process of decomposition of CV structures into independent, re-combinable units.

1. INTRODUCTION

One of the most basic characteristics of modern human language is its particulate structure (cf. e.g. Studdert-Kennedy [21]). In short, lower level items can be re-used and re-combined to produce constructions at higher levels, characterised with new features. At the very basis of this is the compositionality of phonology, where a limited number of basic, meaningless units is used to derive meaningful words.

Modern phonological theories disagree as to what that basic unit might be. Syllables, demisyllables, phonemes, distinctive features, gestures and primes are just some of the proposed candidates. The present paper will try to sketch a scenario of how this compositionality may have come about in the phylogeny of human language, providing some evidence supporting a segment-based approach.

2. FRAMES, THEN CONTENT

An elaborated theory of speech motor control in ontogeny and phylogeny has come from MacNeilage and Davis [3, 12, 13]. Known as the Frame/Content Theory of speech development, it claims that speech develops in infants as a result of superimposition of detailed articulatory content on the basic rhythmic frame derived from mandibular oscillation. Mandibular oscillation is seen as originating in ingestive activity, and claimed to have been evolutionarily co-opted for communicative purposes (with some possible intermediate steps, such as lip-smacks in ancestral hominids). The suggested course of ontogenetic development starts with the combination of cyclic depressions and elevations of the mandible with phonation. Jaw oscillation is claimed to be the only factor responsible for the alternation between vowel-like and consonant-like articulations (corresponding to the open and closed phases, respectively) during vocalisations typical for babbling in human infants. At this stage, variation between consecutive frames comes from shifting of the tongue body mass forwards or backwards; as a result, in addition to the so-called "pure" frames (where the tongue rests in a neutral position), "fronted" and "backed" frames are achieved. The three strategies (pure, fronted and backed frames) are characterised by a strong tendency towards consonant-vowel co-occurrence (labial consonants with central vowels, coronals with front vowels, and dorsals with back vowels), thus typically [baba], [didi] and [gogo] ([3], but see [23] for discussion of possible variation in pure frame shapes based on articulatory modelling).

The next stage corresponds to that of "variegated" babbling in human infants, where the "vocalic" and "consonantal" components become progressively more and more independent: consecutive frames may differ from each other, and the harmony within them gives way to cross-combination of elements. (However, even here MacNeilage and Davis claim that most of the variegation comes from variation in the magnitude of the oscillatory movement, so that vowels vary mainly in height, and consonants – in manner.)

In the phylogenetic timescale, pressure for larger lexicons is claimed to have been the causing factor behind the variegation stage, which evidently made possible the growth of in the numbers of possible CV combinations. Computer simulations (e.g. [18]) have demonstrated that indeed combining the mandibular oscillation principle with pressure for lexicon expansion indeed leads to the development of vocabularies showing a marked preference for CV structures. However, MacNeilage and Davis speak of the CV frame as a "coherent package" based on the strong co-occurrence patterns [3].

3. FROM FRAMES TO COMPOSITIONALITY

There are at least two viable scenarios explaining how those "packages" may have been dissociated into subcomponents, corresponding initially to the close and

ISBN 1-876346-48-5

open cycles of the CV frame, to arrive at the particulate principle responsible for the open-ended digital complexity of modern human phonologies. Either individual gestures were extracted from CV sequences, or multiple gestures were temporally bundled, and such bundles analysed as consistent units.

Lindblom [10, 11] (cf. also Studdert-Kennedy, e.g. [21]) argues that recombination is a natural strategy in a situation where, given a sufficiently large repertoire of utterances, articulatory gestures that tend to re-occur are stored and processed separately due to the principle of minimising the consumption of energy devoted to motor activity. As a result, first the cost of production of those re-occurring components decreases, and novel combinations between them become available.

Goldin-Meadow [9] provides fascinating evidence that strategies of this kind are indeed used by deaf children growing up in families where no sign language is used, and therefore deprived of "normal" language input to guide their development. Such children develop "homesign" systems to satisfy their communication needs, and one of the features of such systems is that complex gestures become decomposed into simpler constituent gestures, which in turn are combined into novel signs.

Computer simulations using populations of autonomous agents forced to imitate "conspecifics" have demonstrated that category formation based on a similar principle is certainly viable in a situation where no shared sound system exists at the outset but members of a population share articulatory and perceptual skills (Oudeyer [15]).

However, even if the articulatory space is constrained anatomically (cf. Lindbom [10, 11]), the re-occurring articulations must be characterised by one important feature: stability. Perkell et al. [16] put forward a speech motor control theory that seems to provide some insight into exactly how such stability may be achieved. In their approach, speech is controlled by an internal model consisting of auditory goals associated with the articulatory configurations used to achieve those goals. This internal model is responsible for fluent speech execution in adult speakers, minimising the need of constant monitoring of the acoustic output during production. Perkell et al. present evidence from normally hearing speakers and subjects deafened after the completion of language acquisition, showing that indeed it is possible to maintain relatively fluent speech production for some time after the loss of auditory feedback but that it does eventually suffer, which shows that (1) there must indeed be some kind of internal model of speech production control that can be used without auditory feedback but at the same time (2) the feedback is necessary to maintain speech production "in good shape".

Perkell et al.'s [16] model has two features that are important for the present proposal. Firstly, the auditory targets correspond to traditionally conceived segments. Secondly, the model introduces the notion of "saturation effects", whereby the auditory feedback gained from speech production is reinforced by orosensory feedback for certain classes of sounds. (mainly consonants but possibly also some vocalic articulations providing relatively stable sensory information, such as [i]).

In line with Perkell et al.'s findings, CV sequences in babbling (and in the pre-speech behaviour of ancestral hominids) may then be suspected to be reinforced through coupling their acoustic effects with the articulatory activity employed to effect them. This is precisely the task of the "Articulatory Filter" proposed by Vihman and DePaolis [22].

4. INTEGRATION

The proposed scenario is thus as follows:

(1) The initial stage in the development of phonological compositionality may have corresponded to CV frames coupled with phonation. The initial variation in place of articulation for the vowel- and consonant-like elements originated from shifts in tongue body position initiated independently of the oscillatory cycle.

(2) This automatically resulted in an increase in the number of possible utterances. However, the range of utterances was still relatively narrow. As a result, frame subcomponents tended to re-occur at sufficiently high frequencies for them to become stored independently. Two important factors here must have been consolidation of the articulatory configurations through "practice" and proprioception, and the cultural development of shared sound repertoires in speaker groups.

(3) After this independence was achieved, the gestural configurations became available for recombination.

5. DISCUSSION

A number of comments are due.

Firstly, it must be borne in mind that the growth of phonological compositionality along the phylogenetic timescale must have been, understandably, a process far more gradual than the development of speech in human infants as we know it today. As is evident from modern non-human primates, ancestral hominids may have possessed some vocal communication system even before the advance towards "variegated proto-babbling". Whether or not the descent of the larynx played a role in making available to them a range of articulations larger than that found in contemporary primates (cf. [7]), it is highly unlikely that the whole range became available at once. This may be seen as actually supporting the present hypothesis, as ample time would have been available for the "practicing" of articulatory options, and the consolidation of the corresponding articulatory-auditory couplings. One additional conclusion here would be that young individuals in such ancestral communities would have at their disposal some acoustic stimuli in their environment. (This might bear some similarity to

ISBN 1-876346-48-5

Bickerton's [1] ideas of the transition from proto-language to language.) With a "lexicon" of a kind already present, it seems doubtful that pressure for lexicon expansion actually drove the growth of compositionality. Rather, in line with Studdert-Kennedy's [21] proposals, vocal imitation and play (perhaps mainly in young individuals against a background of existing adult vocalisations) may have been the driving force. It may have been the growing *availability* of well-grounded, re-usable articulations that made it possible to assign new meanings to newly discovered articulatory combinations.

A number of computer simulations have recently shown that indeed cultural co-operation may be an extremely important factor promoting the growth of relatively stable, shared articulatory behaviours not only in individuals, but across communities (de Boer [4]). There have been simulations that arrived at similar results explicitly denying the role of meaning transmission (Oudeyer [15]), and focussing on imitation and "practice". Note that learning and imitation in a cultural context add to the stability and distinctiveness of the articulations, thus making the search through the articulatory space progressively less random (to paraphrase Lindblom [10]).

Importantly, in such a context, configurations characterised by a greater stability of the association between articulation and acoustics are at a "selective advantage".

The question arises of whether individual gestures may have guided the transition towards compositionally structured phonology. If we assume that the extraction of units fit for recombination (Lindblom [10, 11]) is based on their re-occurrence, which is due to stability and repeatability, and this stability is achieved through articulatory-auditory coupling, then segment-sized, holistic constellations seem to be better candidates than individual gestures. If, as has been shown by Perkell et al. [16], auditory targets are instrumental in forming articulatory-acoustic associations, and if the gestural configurations recruited for re-use must be characterised by articulatory stability, then they must also show stability of the auditory targets within certain limits. This is so because even though the acoustic feedback will contain individual correlates of individual gestures for both consonant- and vowel-like articulations, it will still be holistic. All oral articulations contain at least acoustic imprints of place and manner (or degree of opening and tongue position and lip rounding for vowel-like articulations), in addition to information about the state of the glottis. (See also Gick et al.'s [8] model advocating the role of inter-gestural co-ordination for perceptual "recoverability".)

At the same time, Perkell et al. [16] show that there are "trade-off" relations between individual gestural components of vocalic and sonorant (or possibly even more generally continuant) articulations (that is, the same overall auditory target may be reached using a number of subtly different articulatory strategies), while consonantal articulations – especially plosives – show even greater immunity to articulatory imprecision in certain respects due to the above-mentioned articulatory "saturation effects" (where the auditory feedback is reinforced by sensory feedback), and the inherent elasticity and inertia of the articulators.

It is worth noting that such an approach has one additional advantage: if the articulatory target is indeed holistic in this sense, it will also contain additional information which is not captured by e.g. traditional phonological accounts based on coarse-grained distinctive features. The preservation (through imitation and vocal learning in a cultural context) of such information may be responsible for the fine-grained distinctions between (allegedly very similar) articulations which have become language-specific in the course of language evolution, e.g. the subtle differences between English and German /i/ (cf. Donegan [6]).

Please note, however, that Perkell et al. do not necessarily make a strong claim about the status of the auditory targets relative to their organisation at a higher level – the question of whether e.g. subsequent targets are in any way related is not discussed, the main tenet being their overall stability and purported temporal extent (corresponding to segments).

Davis and MacNeilage [3] suggest that motor control for frames and content is executed using different cerebral structures. Additionally, some evidence has been recently presented for separable processing of consonants and vowels. Caramazza et al. [2] have shown evidence from two aphasic patients whose production of consonants and vowels was affected by their respective conditions to varying degrees, so that errors involved either predominantly consonants or vowels. Monaghan and Shillcock [14] have backed Caramazza et al.'s account with a connectionist model that developed separable processing of consonants and vowels, and suffered from distortion effects reminiscent of those seen in the aphasic patients. Gick et al. [8], Poeppel [17] and others have suggested that this separation may in fact be related to partial lateralisation of the processing of auditory input in the fine- versus coarse grained time domain. The processing of longer temporal windows is claimed to be executed bilaterally, while shorter time windows processing is predominantly lateralised to the left hemisphere.

All of this seems to fit well with classical findings of research into speech errors (e.g. Dell [5], Shattuck-Hufnagel [20]), supported recently with psycholinguistic and computational evidence by Roelofs [19] to the effect that speech planning is performed in terms of segment-sized items "inserted" into rhythmical frames.

6. CONCLUSION

It seems likely, then, that the units into which the initially holistic CV structures were decomposed in the course of evolution of ancestral vocalisations corresponded to segmental units as we know them from modern languages. This argument hinges on the fact that in order to make

 ISBN 1-876346-48-5

possible the extraction of re-occurring, repeatable subcomponents from CV structures, these subcomponents must have been stable. In turn, this stability is seen as stemming from consolidation of production through exploration of one's own articulatory space, and imitation in a cultural of production of conspecifics, equipped with compatible but possibly subtly different, productive mechanisms.

REFERENCES

[1] D. Bickerton, "How protolanguage became language". In C. Knight, M. Studdert-Kennedy and J.R. Hurford (eds.), The evolutionary emergence of language. Cambridge: Cambridge University Press, 2000, pp. 264-284.

[2] A. Caramazza, D. Chialant, R. Capasso, G. Miceli, "Separable processing of consonants and vowels", *Nature*, **403**, pp. 428-430, 2000.

[3] B. Davis and P.F. MacNeilage, "The internal structure of the syllable: An ontogenetic perspective on origins". In T. Givón and B.F. Malle, The evolution of language out of pre-language. Amsterdam/Philadelphia: John Benjamins, 2002, pp. 135-153.

[4] B. de Boer, "Evolving sound systems". In A. Cangelosi and D. Parisi (eds.), Simulating the evolution of language. Berlin: Springer Verlag, 2002, pp. 79-97.

[5] G.S. Dell, "A spreading-activation theory of retrieval in sentence production. *Psychological Review*, **93**, pp. 283-321, 1986.

[6] P. Donegan, "Phonological processes and phonetic rules". In K. Dziubalska-Kołaczyk and J. Weckwerth (eds.), Future challenges for Natural Linguistics. Munich: Lincom, 2002, pp. 57-81.

[7] W.T. Fitch, "Comparative vocal production and the evolution of speech". In A. Wray (ed.), The transition to language. Oxford; Oxford University Press, 2002.

[8] B. Gick, F. Campbell, S. Oh and L. Tamburri-Watt, "Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids". *Journal of Phonetics*, under review.

[9] S. Goldin-Meadow, "Getting a handle on language creation". In T. Givón and B.F. Malle, The evolution of language out of pre-language. Amsterdam/Philadelphia: John Benjamins, pp.343-374, 2002.

[10] B. Lindblom, "Systemic constraints and adaptive change in the formation of sound structure". In J.R. Hurford, M. Studdert-Kennedy and C. Knight (eds.), Approaches to the evolution of language. Cambridge: Cambridge University Press, 1998, pp. 242-264.

[11] B. Lindblom, "Developmental origins of adult phonology: The interplay between phonetic emergents and the evolutionary adaptations of sound patterns". *Phonetica* **57**(2-4), pp. 297-314, 2000.

[12] P.F. MacNeilage, "The frame/content theory of evolution of speech production", *Behavioral and Brain Sciences* **21**, pp. 499-511, 1998.

[13] P.F. MacNeilage and B. Davis, "On the origins of intersyllabic complexity". In T. Givón and B.F. Malle, The evolution of language out of pre-language. Amsterdam/Philadelphia: John Benjamins, 2002, pp. 155-170.

[14] P. Monaghan and R. Shillcock, "Connectionist modelling of the separable processing of consonants and vowels", *Brain and Language*, in press.

[15] P.-y. Oudeyer, "Phonemic coding might be a result of sensory-motor coupling dynamics". In B. Hallam, D. Floreano, J. Hallam, G. Hayes, J-A. Meyer (eds.), *Proceedings of the 7th International Conference on the Simulation of Adaptive Behavior*, MIT Press, 2002, pp. 406-416.

[16] J.S. Perkell, F.H. Guenther, H. Lane, M.L. Matthies, P. Perrier, J. Vick, R. Wilhelms-Tricarico and M. Zandipour, "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss", *Journal of Phonetics*, **28**, pp.233-272, 2000.

[17] D. Poeppel, "Pure word deafness and the bilateral processing of the speech code". *Cognitive Science*, **25**, pp. 697-693, 2001.

[18] M.A. Redford, C.C. Chen and R. Miikkulainen, "Constrained emergence of universals and variation in syllable systems", *Language and Speech*, **44**, pp. 28-57, 2001.

[19] A. Roelofs, "Phonological segments and features as planning units in speech production". *Language and Cognitive Processes*, **14**(2), pp. 173-200, 1999.

[20] S. Shattuck-Hufnagel, "Speech errors as evidence for a serial-order mechanism in sentence production". In W.E. Cooper and E.C.T. Walker (eds.), Sentence processing: Psycholinguistic studies presented to Merrill Garrett. Hillsdale, NJ: Lawrence Erlbaum, 1979, pp. 295-342.

[21] M. Studdert-Kennedy, "Evolutionary implications of the particulate principle: Imitation and the dissociation of phonetic form from semantic function". In C. Knight et al. (eds.), pp. 161-176.

[22] M.M. Vihman and R.A. DePaolis, "The role of mimesis in infant language development: evidence for phylogeny?" In: C. Knight, M. Studdert-Kennedy and J.R. Hurford (eds.), The evolutionary emergence of language. Cambridge: Cambridge University Press, 2000, pp. 130-145.

[23] A. Vilain, C. Abry, P. Badin and S. Brosda, "From idiosyncratic pure frames to variegated babbling: Evidence from articulatory modelling". *Proceedings of the 14th IcphS*, San Francisco, pp. 2497-2500, 1999.

ISBN 1-876346-48-5