Exploring the sound structure of novel vocalizations

Susanne Fuchs[*1], Šárka Kadavá[*1,2,3], Wim Pouw[3], Bradley Walker[4], Nicolas Fay[4], Bodo Winter[5], and Aleksandra Ćwiek[1]

[*]Shared first authorship, corresponding authors: fuchs|kadava@leibniz-zas.de
[1]Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany
[2]Linguistik, Georg-August Universität, Göttingen, Germany
[3]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
[4]School of Psychological Science, The University of Western Australia, Perth, Australia
[5]Depart. of English Language and Linguistics, Uni of Birmingham, Birmingham, United Kingdom

When humans speak or animals vocalize, they can produce sounds that are further combined into larger sequences. The flexibility of sound combinations into larger meaningful sequences is one of the hallmarks of human language. To some extent, this has also been found in other species, like chimpanzees and birds. The current study investigates the structure of sounds when speakers are asked to communicate the meaning of 20 selected concepts without using language. Our results show that the structure of sounds between pauses is frequently limited to 1–3 sounds. This structure is less complex than when humans use their native language. The acoustic distance between sounds depends largely on the concept apart from concepts referring to animals, which show a higher diversity of involved sounds. This exploratory analysis might provide evidence of how the structure of sound could have changed from simple to complex in evolution.

1. Introduction

Human speech is composed of small units: sounds that are meaning-distinguishing (phonemes). Several sounds combine into syllables, words, and phrases that carry meaning(s). The sequential order of sounds into larger sequences is a milestone in speech acquisition, and already young infants can start producing sequences of vocalization before they acquire their mother tongue (Wermke, Robb, & Schluter, 2021). Even when language is acquired, nonverbal vocalizations are present in adult communication and are an emerging field of study at the boundaries between non-human and human communication (Pisanski, Bryant, Cornec, Anikin, & Reby, 2022). That means sequences of sounds are not a property of human communication alone but are also found in non-human animals like birds (Sainburg, Theilman, Thielk, & Gentner, 2019; Doupe & Kuhl, 1999; Favaro et al., 2020), meerkats (Rauber, Kranstauber, & Manser, 2020), chimpanzees (Girard-Buttoz et al., 2022). Comparative approaches between human and non-human animal vocalization deserve bottom-up methodologies rather than human-centric analyses (Hoeschele, Wagner, & Mann, 2023). What has been

called a syllable in non-human vocalization refers to sound(s) produced between pauses. In human speech production, similar chunks have often been termed interpausal units (Bigi & Priego-Valverde, 2019; Prakash & Murthy, 2019). They refer to speech that is realized between pauses.

In this exploratory study, we are interested in sounds realized in novel vocalizations during a charade game, i.e., in a situation where the use of actual words of the participant's language is 'forbidden'. This paradigm has been used to investigate the origin and evolution of language (Fay et al., 2022; Ćwiek et al., 2021; Perlman & Lupyan, 2018).

This paper aims to explore how many sounds are realized between pauses in non-linguistic vocalizations. Furthermore, we investigate the diversity of sounds realized within different concepts, by assessing the distance between them in a multi-variable acoustic space.

2. Methodology

2.1. *Corpus creation*

The present study uses a subset of data collected in a larger study in which participants were recorded performing a series of concepts in three conditions. In the three conditions, participants are asked to communicate a set of concepts using either (1) only gestures, (2) only non-linguistic vocalizations and other sounds, or (3) a combination of gestures and vocalizations. Here, we focus on a subset of the vocalization recordings. We have not analyzed the vocalizations that are produced in the multimodal condition because we assume that first, they are not stand-alone carriers of the meaning, and second, their forms are shaped by the coordination with body motion.

The recordings analyzed here were produced by 62 first-year psychology students at the University of Western Australia (43 female, 17 male, 2 non-binary; aged 17–33, M = 20.21, SD = 3.36). All were speakers of English. Of these, 28 participated in person and 34 remotely via Microsoft Teams, due to COVID-19 restrictions. Participants were allocated 60 concepts to communicate (20 in each modality condition), sampled from a list of 200 concepts comprising the 100-item Leipzig-Jakarta list of basic vocabulary (Tadmor, 2009) plus 100 other basic concepts chosen based on their sensory and modality preferences (Lynott, Connell, Brysbaert, Brand, & Carney, 2020). They were asked to communicate each concept using the specified modality (and without using language) so that another person would be able to view the recording and guess the concept from a list of options. If the participants could not think of a way to communicate a concept, they were permitted to skip it.

2.2. *Concept extraction*

For the exploratory analysis, we focused on a variety of concepts that might reflect different degrees of concreteness and abstraction (see 1). For example, the concept *maybe* is rather abstract or logical than *smoke*. We chose these different concepts to have a wider semantic potential, but have not added categories to the concepts, because a dichotomy between concreteness vs. abstraction has currently been questioned (Banks et al., 2023).

Our analysis only included concepts for which initially at least 5 participants produced vocalizations. For three concepts we excluded acoustic trials as they contained a considerable amount of background noise that made an analysis unreliable.

Table 1. Concepts used in this study. L-J corresponds to the Leipzig-Jakarta list.

Concept	List	No. of speakers
happy	other	6
sad	other	7
bad	other	7
scared	other	5
good	L-J	6
angry	other	7
disgusted	other	7
dog	L-J	6
cat	other	6
bird	L-J	5
fish	L-J	5
fly	L-J	8
old	L-J	4
spoon	other	5
egg	L-J	6
ash	L-J	3
stone/rock	L-J	6
smoke	L-J	4
maybe	other	8
not	L-J	7

2.3. *Acoustic annotation procedures*

The acoustic data were labeled in Praat 6.1.51 (Boersma & Weenink, 2021) by three annotators who are phoneticians by training. Following Swets, Fuchs, Krivokapić, and Petrone (2021), all silent intervals longer than 100 ms were treated as pauses and labeled with 'p'. Apart from placing boundaries next to pauses, the annotators additionally labeled successive sounds without pauses. The following criteria were used in the decision-making process for separating the speech stream into two or more sounds: a) two (or more) prominent amplitude peaks in the amplitude envelope were present, b) changes in spectral characteristics (e.g., formant structures) were present, and c) sounds were perceptually distinct. Variations in fundamental frequency, e.g., a downward and then upward motion, were only considered as two sounds when they also showed spectral differences in higher frequency ranges and/or differences in the amplitude envelope. All sounds were labeled with an initial 's' and successive numbers when they occurred in a sequence. The first annotator (a1) created the annotation criteria and labeled the data. Annotator 2 (a2) used the available TextGrids from a1 and changed the boundaries when she disagreed. Both agreed on 94.6 percent of the number of sounds. Hereafter, a1 inspected all acoustic files again where disagreement was found and confirmed the

changes. Annotator 3 (a3) started labeling from scratch without having TextGrids available. Inter-rater agreement between a2 and a3 was 96.7 percent concerning the overall number of labeled sounds. The temporal differences between the onset of a given sound labeled by a2 and its closed temporal neighbor labeled by a3 were calculated. The same was done for the offset of a sound. The differences were on average 0.048s (median = 0.018s) for the onset and 0.088s (median = 0.027s) for the offset. These differences are influenced by the number of sounds an annotator labeled for a given concept, which makes the calculation of inter-rater agreement challenging. We think that for the current exploratory analysis, the overall agreement is reasonable. We decided to take a2's segmentation for further analysis.

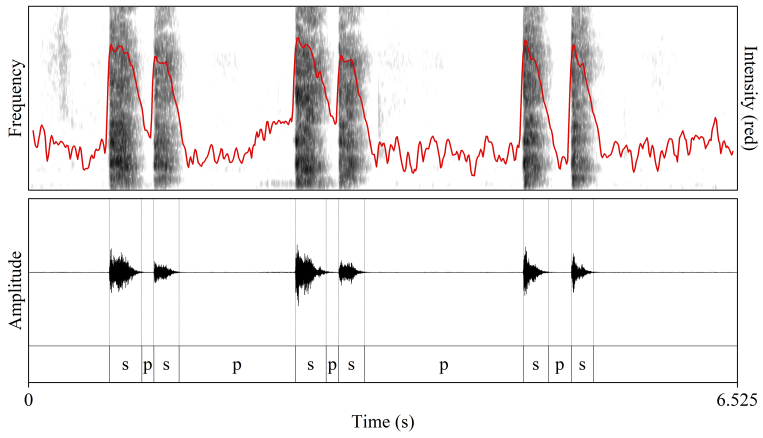

Figure 1. Example for acoustic annotation of the concept *smoke*. All segments are labeled as 's' and pauses as 'p'. The red line depicts the intensity curve.

2.4. *Analyses of acoustic similarity*

Initially, all audio files were cut into segment-sized files using a custom Python script. Acoustic analysis was performed on these sounds, using the analyze() function of the `soundgen` package in `R` (Anikin, 2019). The output of this function consists of more than 100 acoustic parameters as listed in the documentation (e.g., $f0$, amplitude, formant values, entropy, and their respective mean, median, and standard deviation). Some of these acoustic parameters are present or absent in the recorded sounds, e.g., voicing. However, the presence of voicing is redundant with intensity because voiced sounds are louder than voiceless ones and intensity

values can always be calculated. That means, some acoustic parameters are highly correlated and redundant with others. For this reason, we excluded parameters resulting in NA values in the post-processing. Moreover, we excluded voice quality parameters (e.g., flux), because these parameters may have been very sensitive to background noise, which occurred in some speakers. All final parameters were averaged for the whole time series of a sound, and we used mean and standard deviation for further explorations. We ended up with a multidimensional dataset consisting of 45 acoustic parameters. For the analysis of acoustic similarity, we calculated the Euclidean distance between the vector of acoustic parameters of each sound, to all other sounds. As a result, we got a distance matrix that allowed us to extract an average distance between sounds within a trial of a concept and compare it to other concepts.

3. Results and Discussion

3.1. *Structural similarity*

To explore structural similarity, we analyzed if certain sounds occurring between pauses appear alone or in successive order. When speakers try to communicate concepts using novel vocalizations, they frequently realize a relatively small number of sounds between two pauses: 1 sound occurred 208 times, 2 sounds = 80 times, 3 sounds = 35 times, 4 sounds = 24 times, 5 sounds = 11 times, 6 sounds = 3 times, 8 sounds = 4 times, 9 sounds = 1 time, 10 sounds = 1 time, 16 sounds = 2 times, 18 sounds = 1 time. That means structurally most concepts (208 cases in our dataset) are realized with only one sound <s> that is surrounded by pauses. In 80 cases we found realizations of two successive sounds <ss> and in 35 cases participants produced three successive sounds <sss> without being interrupted by a pause. If the data are split by concept, vocalizations for *cat*, *dog*, and *bird* (all within a broader class of animals) also have more than three successive sound combinations, probably mirroring onomatopoeia. For the rest of the data, no conclusions can be drawn, because the number of sounds between pauses is concept-specific.

If pauses are taken into account, sounds were combined flexibly, for example, for four sounds we could get combinations such as <s|s|s|s> or <ss|ss> or <ss|s|s> where | marks a pause.

3.2. *Acoustic similarity*

Similar sounds may be repeated, like in imitating 'coo-coo', or they may be of different acoustic quality, like in imitating a cat's 'meow'. For this reason, we were further interested in examining the similarity between sounds that make up a novel vocalization.

To have a first look into the diversity of sounds, we analyzed their average acoustic distance within each trial. We preferred this data-driven approach in

contrast to labeling the data to phonemic features because it allows us to include sounds that may not occur in the English phoneme inventory, e.g., whistles or clicks. It represents continuous acoustic data instead of putting categorical labels to it, which could also be biased by the native language of the annotator.

Figure 2 depicts the results. We can see that the different concepts vary in their average acoustic distance between sounds. Some abstract concepts like *not* consist of sounds that are closer to each other in distance (i.e., more similar), while *dog* has a larger average acoustic distance between the sounds. Those concepts with several successive sounds (e.g., <sss>) are also the ones with the largest average distance.

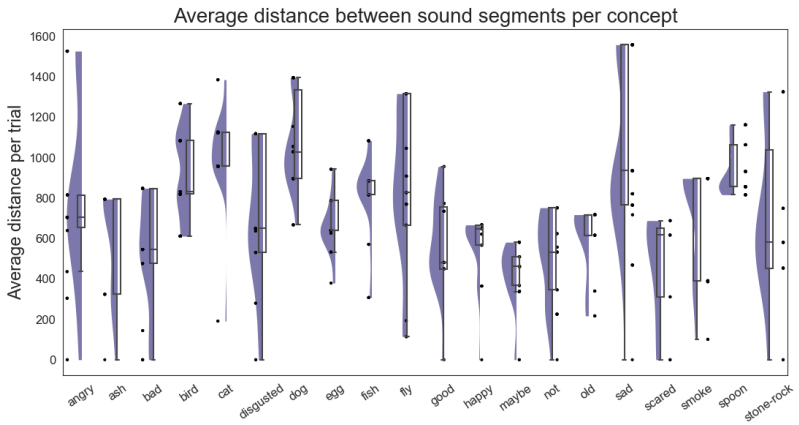

Figure 2. Average acoustic distances between sounds within a single trial displayed by concept, boxplots, and half-violins in purple display data distribution, black dots correspond to single trials. Each concept is displayed at the x-axis and ordered by alphabet.

In summary, the structure of novel vocalizations obtained from a charade game most often contains either one, two, or three successive sounds that are not separated by pauses. This may to some extent be similar to infant's vocalization (Wermke et al., 2021) and non-human species. It is different from human speech production, where already syllables or morphemes can consist of three sounds. Those are combined into larger chunks that are not interrupted by pauses. Our findings suggest that novel vocalizations have a rather simple sound structure that is complexified (i.e., more and probably shorter sounds are realized in a sequence) during language evolution.

4. Supplementary Materials

Dataset and scripts are available on `https://github.com/sarkadava/Evolang2024_SoundSimilarity`.

Acknowledgements

We like to thank the reviewers of Evolang, the participants of the study, and Melissa Ebert for data annotation. This work has been supported by a grant from the German Research Council (FU791/9-1).

References

Anikin, A. (2019). Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior research methods*, *51*, 778–792.

Banks, B., Borghi, A. M., Fargier, R., Fini, C., Jonauskaite, D., Mazzuca, C., Montalti, M., Villani, C., & Woodin, G. (2023). Consensus paper: Current perspectives on abstract concepts and future research directions. *Journal of Cognition*, *6*(1).

Bigi, B., & Priego-Valverde, B. (2019). Search for inter-pausal units: application to cheese! corpus. In *9th language & technology conference: Human language technologies as a challenge for computer science and linguistics* (pp. 289–293).

Boersma, P., & Weenink, D. (2021). Praat: doing phonetics by computer [computer program](2011). *Version*, *5*(3), 74.

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., et al.. (2021). Novel vocalizations are understood across cultures. *Scientific Reports*, *11*(1), 10108.

Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience*, *22*(1), 567–631.

Favaro, L., Gamba, M., Cresta, E., Fumagalli, E., Bandoli, F., Pilenga, C., Isaja, V., Mathevon, N., & Reby, D. (2020). Do penguins' vocal sequences conform to linguistic laws? *Biology letters*, *16*(2), 20190589.

Fay, N., Walker, B., Ellison, T. M., Blundell, Z., De Kleine, N., Garde, M., Lister, C. J., & Goldin-Meadow, S. (2022). Gesture is the primary modality for language creation. *Proceedings of the Royal Society B*, *289*(1970), 20220066.

Girard-Buttoz, C., Zaccarella, E., Bortolato, T., Friederici, A. D., Wittig, R. M., & Crockford, C. (2022). Chimpanzees produce diverse vocal sequences with ordered and recombinatorial properties. *Communications Biology*, *5*(1), 410.

Hoeschele, M., Wagner, B., & Mann, D. C. (2023). Lessons learned in animal acoustic cognition through comparisons with humans. *Animal Cognition*, *26*(1), 97–116.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, *52*, 1271–1291.

Perlman, M., & Lupyan, G. (2018). People can create iconic vocalizations to

communicate various meanings to naïve listeners. *Scientific reports*, *8*(1), 2634.

Pisanski, K., Bryant, G. A., Cornec, C., Anikin, A., & Reby, D. (2022). Form follows function in human nonverbal vocalisations. *Ethology Ecology & Evolution*, *34*(3), 303–321.

Prakash, J. J., & Murthy, H. A. (2019). Analysis of inter-pausal units in indian languages and its application to text-to-speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(10), 1616–1628.

Rauber, R., Kranstauber, B., & Manser, M. B. (2020). Call order within vocal sequences of meerkats contains temporary contextual and individual information. *BMC biology*, *18*, 1–11.

Sainburg, T., Theilman, B., Thielk, M., & Gentner, T. Q. (2019). Parallels in the sequential organization of birdsong and human speech. *Nature communications*, *10*(1), 3636.

Swets, B., Fuchs, S., Krivokapić, J., & Petrone, C. (2021). A cross-linguistic study of individual differences in speech planning. *Frontiers in Psychology*, *12*, 655516.

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. *Loanwords in the world's languages: A comparative handbook*, *55*, 75.

Wermke, K., Robb, M. P., & Schluter, P. J. (2021). Melody complexity of infants' cry and non-cry vocalisations increases across the first six months. *Scientific reports*, *11*(1), 4137.