The learnability of the Nepali four way stop-voicing contrast: Three categories are learnable despite hypo-articulation, but the fourth is adrift

Titia Benders, Katherine Demuth

Department of Linguistics, Macquarie University
Titia.Benders@mq.edu.au, Katherine.Demuth@mq.edu.au

ABSTRACT

Can infants learn consonant categories from hypo-articulated infant-directed speech? This question was addressed with learnability analyses on a previously published corpus of the Nepali four-way voicing contrast [4]. Both a supervised (Discriminant Analysis) and an unsupervised (Gaussian Mixture Model) learner successfully learned the voiceless-aspirated, voiceless-unaspirated, and voiced-unaspirated stop. Neither mechanism could learn the voiced-aspirated category from the lead- and lag-time distributions. Implications are discussed for infants' acquisition of the Nepali four-way contrast, language change, and their relationship.

Keywords: Infant-directed speech; Learnability; Nepali; Stop-voicing contrast

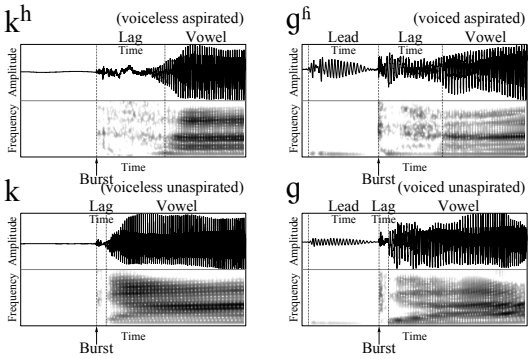
1. INTRODUCTION

The acoustic speech sound distributions in infants' input undisputedly play a crucial role in infants' perceptual attunement to native-language consonant and vowel contrast [25; 20]. Attunement may be aided by hyper-articulated input, i.e., enhanced mean distances between the acoustic categories [17]. However, not all parents hyper-articulate, with hypo-articulation being frequently attested as well [3; 8; 18]. This raises questions about the learnability of hypo-articulated input.

A recently attested case of hypo-articulated input to infants is the Nepali four-way voicing contrast [3]. Nepali, like many other Indo-Aryan languages, canonically contrasts four voicing categories for every obstruent place of articulation [1] (Fig. 1). Such four-way contrasts cannot be described using voice-onset time. Rather, the voiced-voiceless contrast is canonically cued by the presence versus absence of lead time - the onset of vocal fold vibration before the burst; the aspirated-unaspirated contrast is canonically cued by the duration of lag time - the delay between the burst and the onset of the vowel [17] (Fig. 1). Nepali mothers speaking to their infants in so-called infant-directed speech (IDS) reduce the lead-time contrast between voiced and voiceless stops and shorten the lag time of all their stops [3]. The present study addresses the learnability ramifications of this hypo-articulation.

A complicating issue with respect to the Nepali four-way voicing contrast is the potential instability

Figure 1: Waveforms (top), spectrograms (bottom), and annotation of canonical tokens of the Nepali four-way stop-voicing contrast.



of the voiced-aspirated obstruents. A consideration of other Indo-Aryan languages finds, for example, that the voiced-aspirated stop has disappeared from Kashmiri [4] and has been replaced by a tonal contrast in Punjabi [13; 15]. Within Nepali, the aspiration of voiced-aspirated obstruents is variably produced for stops in intervocalic and word-final position [17], and completely replaced by a lowered F0 and increased breathiness for affricates [5]. Yet, Nepali is said to still have the canonical four-way contrast word-initially [17]. The present study will assess whether the lead- and lag-time distributions from hypo-articulated IDS support the acquisition of all four stop-voicing categories in highly common words with velars.

The learnability of the Nepali four-way voicing contrast will be addressed in two types of computational learnability analyses, corresponding to two broad hypotheses about infants' perceptual attunement. The first hypothesis is that infants have access to bottom-up acoustic as well as top-down lexical information to guide their perceptual attunement [11; 23]. This supervised learning hypothesis will be implemented using a Discriminant Analysis (DA), which is provided with the input tokens' lead- and lag times as well as their category membership. The second hypothesis is that infants only have access to bottom-up acoustic information [19]. This unsupervised learning hypothesis will be implemented using Gaussian Mixture Models (GMMs), which estimate the multivariate Gaussian category structure underlying the observed continuous distributions. The results from both sets of analyses will answer the question to what extent the

Nepali four-way voicing contrast is learnable from the lead- and lag-time distributions in Nepali mothers' hypo-articulated IDS.

2. METHODS

2.1. Corpus

The present study uses a corpus first presented in [3]. Here we repeat key information for interpreting the present results, and minor deviations from [3] in token selection.

2.1.1. Participants & Procedure

Participants were 16 female native speakers of Nepali and their infant (aged 10-18 months). Mothers were born and lived at least until puberty in Nepal. Dyads lived in Sydney where the recordings took place.

Mother-infant dyads participated in a play session during which the mother spoke to her infant in IDS about pictures of a *hairpin* /ka.ʈa/, *meal* /kʰa.na/, *bullock cart* /ga.ɖa/, and *neck* /gʱa.ʈi/.

2.1.2. Acoustic measurements and token selection

Tokens of the four target words produced in either isolation or utterance-initial position were manually annotated for lead time (from the onset of prevoicing to the burst) and lag time (from the burst to the onset of clear F2 in the vowel; Fig. 1). Tokens were excluded if they contained external sound overlap, an atypical voice quality, or case marking. [3] then excluded any statistical outliers, to meet their analysis assumptions. The present study included those tokens, as outliers were confirmed to be correctly measured and thus form part of the input.

The analyses reported here include the 935 tokens produced in IDS (/kʰ/=297; /k/=282; /gʱ/=160; /g/=196). All analyses are performed on the natural logarithms of lead and lag time in milliseconds. As the logarithmic transformation of 0 lead times rendered impossible values, these were manually re-set to 0. For the visualizations, jitter around 0 has been added.

2.2. Discriminant analysis

A linear DA was performed using the *lda* function in the *MASS* package [24] of *R* [22]. The four stops were the classes, and the scaled and centred natural logarithms of lead and lag time were the continuous predictors. A DA computes discriminant functions, to maximize the ratio of the variation between and within the classes across the predictors. Discriminant scores per token can be computed and converted into DA-predicted class memberships. These can be compared to the actual targets using a confusion matrix, to assess model success.

2.3. Gaussian Mixture Modelling

The parameters of the multivariate Gaussian distributions that are most likely to have generated the lead- and lag-time were estimated using the Expectation-Maximization algorithm [7], using the *Mclust* function in the *mclust* package [12] of *R* [22]. The parameters to be estimated for each multivariate distribution in these GMMs are the means, covariances, and mixing proportions. A further GMM parameter is the number of categories. Although language-learning infants need to infer this number from their input, learnability simulations with EM GMM typically pre-set the number of categories [2; 6]. For reasons to become apparent when the DA results are presented, GMM analyses were conducted for four as well as three categories. Model comparisons with the Bayesian Information Criterion [22] were employed to establish whether the model with three or four categories is to be preferred, and thus whether the input is compatible with Nepali-learning infants acquiring three or four stop-voicing categories.

The GMM parameters, once estimated, can be used to assign corpus tokens to the acquired categories, which can be labelled after the most prevalent target category among the tokens categorized into a GMM category. Confusion matrices comparing tokens' target and GMM-assigned category memberships were used to evaluate model success.

3. RESULTS

3.1. Visual data inspection

Fig. 2 displays the distribution of IDS tokens across lead and lag time. The canonical description of the Nepali four-way voicing contrast led us to expect four distinct clusters of tokens: the three ellipses in Fig. 2 and a fourth cluster of voiced aspirated tokens produced with lead-time as well as long lag time. However, only a few voiced-aspirated tokens occupy this area of the acoustic space, with the remaining tokens broadly falling within one of the other three categories.

3.2. Discriminant Analysis

The first discriminant function of the DA captured 76.9% of the between-class variance, with coefficients for lead time (-1.244) and lag time (1.165) having opposite signs but almost equal sizes. The DA was at least 89% correct in classifying the target voiceless-aspirated, voiceless-unaspirated, and voiced-unaspirated tokens (Table 2), confirming their visually observed separability in the lead-time/lag-time space. In contrast, the DA was only 1% correct in categorising the target voiced-aspirated tokens, and

Figure 2: Tokens in lead- and lag time acoustic space. Ellipses indicate 95% normal confidence intervals around voiceless aspirated, voiceless unaspirated, and voiced unaspirated tokens.

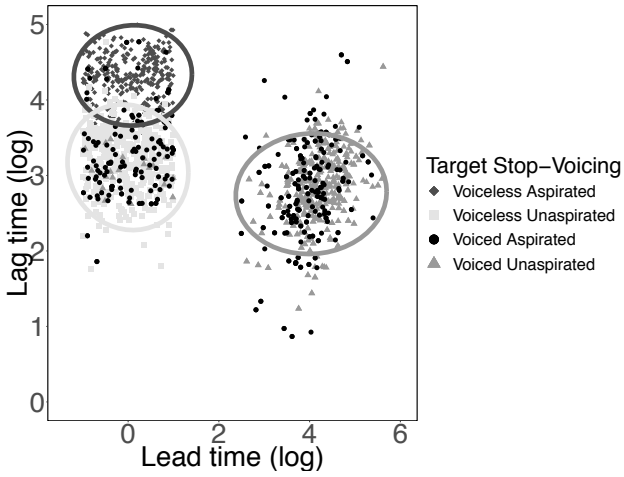


Figure 3: Category memberships as assigned by the 3-category GMM (grey-scale coded, see legend). Target stop-voicing is shape coded (see Fig. 2 legend).

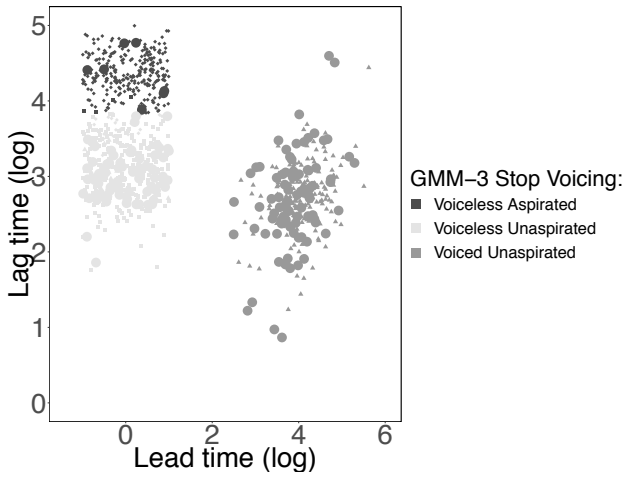


Table 1: Confusion matrix of the Target categories and the DA-assigned categories.

		DA-assigned category			
		kʰ	k	gʱ	g
Target	kʰ	95%	5%	0%	0%
	k	4%	96%	0%	1%
	gʱ	3%	41%	1%	52%
	g	0%	10%	2%	89%

Table 2: Confusion matrix of the Target categories and the 3-category GMM-assigned categories.

		GMM-assigned category		
		kʰ	k	g
Target	kʰ	92%	8%	0%
	k	1%	99%	0%
	gʱ	4%	43%	53%
	g	0%	10%	90%

largely assigned them to the voiceless and voiced unaspirated categories.

3.3. Gaussian Mixture Modelling

Since Nepali is described as having a four-way stop-voicing contrast, whereas the visual inspection and DA analyses suggest a 3-way contrast, two GMMs were fit: one with four, and the other with three categories. Model comparison indicated that the 3-category model (BIC=-2691.114) was preferred over the 4-category model (BIC=-2352.859).

The 3-category GMM (displayed in Fig. 3) and the 4-category GMM were nearly identical in terms of their voiceless-aspirated and voiced-unaspirated categories. The models differed in that the 3-category model detected one voiceless-unaspirated category, whereas the 4-category model detected two categories in that region. Critically, neither model detected the voiced-aspirated category.

The three categories learned by the 3-category model were highly consistent with the target categories, as shown by the over 90% correct classifications of the voiced-unaspirated, voiceless- unaspirated, and voiced-aspirated tokens. The 4-category model performed considerably worse only for the 'split' voiceless-unaspirated category (results not displayed due to space constraints). Both models distributed the voiced-aspirated targets across their voiceless- and voiced-unaspirated categories.

4. DISCUSSION

The present study used computational analyses to investigate whether Nepali-learning infants could acquire the four-way voicing contrast from the lead- and lag-time distributions in Nepali IDS, which is hypo-articulated [3]. Both supervised (DA) and unsupervised (GMM) learning simulations found that Nepali IDS supports the acquisition of three, but not the fourth (voiced-aspirated) category.

The apparent reduction of the Nepali four-way system to a three-way contrast between voiced-unaspirated, voiceless-unaspirated, and voiceless-aspirated stops, conforms to the change observed in several other Indo-Aryan languages [4; 13]. And although some instability of the Nepali voiced-aspirated obstruents has been documented previously [5; 17], the present results are to our knowledge the first suggestion that word-initial stops may no longer be realized with canonical prevoicing and aspiration.

The voiced-aspirated stops in this study were all elicited utterance-initially in a single word (neck-/gʱa.ʈi/), yet variably realized with the lead- and lag time properties of all three other categories. This suggests that speakers of Nepali may still have a

somewhat distinct representation of the voiced-aspirated stop. Further research is needed to establish whether or not listeners reliably identify the voiced-aspirated stop, which would provide evidence for the presence of other cues. Primary cues to consider are the breathiness and F0 of the following vowel, which signal voiced-aspirated affricates in Nepali and stops in Punjabi, respectively [5, 13]. And if a new cue to the voiced-aspirated stop is found, its impact on the learnability of the Nepali four-way obstruent series would need to be considered.

The other three categories were successfully learned by the computational models, despite the hypo-articulated acoustic contrast between voiced and voiceless stops in Nepali IDS [3]. Previous computational work had suggested that the learnability of segmental contrasts is improved by IDS hyper-articulation [6] or clearly articulated tokens under focus [2]. The present findings, resonating the conclusions from [9], show that hypo-articulation doesn't always equate with low learnability.

These three learnable stops were successfully acquired by the supervised (DA) as well as the unsupervised (GMM) model. These findings could be taken as support for bottom-up theories of perceptual attunement [19], suggesting that top-down information may be more important for the acquisition of vowels [11] rather than consonants. However, our models were only provided with tokens of a single word, which could be considered as a form of top-down information. Moreover, the unsupervised learning model was provided with the correct number of categories, while real infants have to infer that number from the input. When our GMMs were required to estimate the number of categories from the data, they found up to nine categories, possibly due to their strict parametric assumptions. Simulations with data from multiple words using non-parametric models will be required to fully explore the power of unsupervised learning for perceptual attunement to consonants.

The learnability of three stop-voicing categories in Nepali, combined with the unlearnable voiced-aspirated stop, raises critical questions about how real infants and children acquire the Nepali voicing system. If unsupervised learning from the acoustic input distributions lays the foundation for infants' early phonological systems [19], Nepali-learning infants may only acquire three categories. If the lexicon is involved in perceptual attunement [11; 23], the free variation displayed by the voiced-aspirated stops may be sufficient for infants to also acquire this fourth category. However, one could also hypothesize that infants and children regularize variation in their language input [14] and are thus instrumental to the disappearance of the voiced-aspirated stops.

The extent to which the voiced-aspirated stop will disappear from Nepali possibly also depends on the impact of orthography on first-language phonology. The voiced aspirated stops are represented with separate graphemes in the Devanagari script that is used for Nepali. Orthography can be a source of lexical context for second-language learners [10], and literate native speakers of Nepali sometimes introduce spelling-based contrasts in their pronunciations [17]. Both perception and production research with Nepali-learning infants and children before, during, and after the onset of literacy is thus required to assess the acquisition of the four-way voicing system, and the impact of first-language learning mechanisms and orthography on the disappearance and maintenance of phonological contrasts.

In any further research on the Nepali obstruent voicing, several limitations of the present study will need to be overcome. Firstly, only velar stops were elicited using only one word per segment. Moreover, the speakers in the present study were part of the Nepali diaspora at the time of the recordings. Future research thus needs to establish to what extent the present findings generalize across words, to other places of articulation, and to speakers still residing in Nepal. Further extensions to this work would also need to consider the frequency and functional load of the voiced-aspirated stop – data that are currently difficult to provide due to limited availability of searchable dictionaries and corpora of Nepali. Such work would shed important light on early phonological acquisition in the face of unstable phonetic realizations.

5. ACKNOWLEDGEMENTS

The authors gratefully acknowledge Sujal Pokharel, who was instrumental in corpus development. This work was funded, in part, by the following grants: ARC FL130100014, ARC Centre for Cognition and its Disorders ARC CE110001021. Equipment was funded by MQ SIS9201501719. We thank Elaine Schmidt, Ivan Yuen, Rebecca Holt, Ping Tang, and members of the Child Language Lab and Phonetics Lab at Macquarie University for helpful comments, and Mrs. Mira Neupane for recruitment assistance.

6. REFERENCES

[1] Acharya, J. (1991). *A descriptive grammar of Nepali and an analyzed corpus*. Georgetown University Press.

[2] Adriaans, F., Swingley, D. (2017). Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *The Journal of the Acoustical Society of America* 141(5), 3070–3078. https://doi.org/10.1121/1.4982246

[3] Benders, T., Pokharel, S., Demuth, K. Hypo-articulation of the four-way voicing contrast in Nepali

infant-directed speech. *Language Learning and Development*, DOI: 10.1080/15475441.2019.1577139

[4] Cardona, G., Luraghi, S. (2018). Indo-Aryan Languages. In B. Comrie (Ed.), *The world's major languages* (2nd ed., pp. 373–379).

[5] Clements, G. N., Khatiwada, R. (2007). Phonetic Realization of Contrastively Aspirated Affricates in Nepali. *Proc. 16th ICPhS* Saarbrücken, 629–632.

[6] De Boer, B., Kuhl, P. K. (2003). Investigating the Role of Infant-Directed Speech with a Computer Model. *Acoustics Research Letters Online* 4(4), 129–134.

[7] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. *Royal Stat. Soc. B* 39, 1–38

[8] Englund, K. T. (2018). Hypoarticulation in infant-directed speech. *Applied Psycholinguistics* 39, 67–87.

[9] Eaves, B. S., Feldman, N. H., Griffiths, T. L., Shafto, P. (2016). Infant-directed speech is consistent with teaching. *Psychological Review,* 123(6), 758–771. https://doi.org/10.1037/rev0000031

[10] Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics* 36(2), 345–360. https://doi.org/10.1016/j.wocn.2007.11.002

[11] Feldman, N. H., Griffiths, T. L., Goldwater, S., Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review* 120(4), 751–778. https://doi.org/10.1037/a0034245

[12] Fraley, C., and Raftery, A. E. (2006). MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering, Technical Report (University of Washington, Seattle, WA).

[13] Gill, H. S., Gleason, H. A. (1969). *A reference grammar of Punjabi*. Department of Linguistics, Punjabi University Patiala.

[14] Hudson-Kam, C. L., Newport, E. (2005). Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change. *Language Learning and Development* 1(2), 151–195. https://doi.org/10.1207/s15473341lld0102_3

[15] Kanwal, J., Ritchart, A. (2015). An Experimental Investigation of tonogenesis in Punjabi. *Proc 18th ICPhS*. https://doi.org/10.1115/1.2898876

[16] Khatiwada, R. (2009). Nepali. *Journal of the International Phonetic Association* 39(3), 373–380. https://doi.org/10.1017/S0025100309990181

[17] Liu, H. M., Kuhl, P. K., Tsao, F. M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, 6(3), 1–10. https://doi.org/10.1111/1467-7687.00275

[18] Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., Cristia, A. (2015). Mothers Speak Less Clearly to Infants Than to Adults: A Comprehensive Test of the Hyperarticulation Hypothesis. *Psychological Science* 26(3), 341–347. https://doi.org/10.1177/0956797614562453

[19] Maye, J., Werker, J. F., Gerken, L. A. (2002). Infant Sensitivity to Distributional Information can Affect Phonetic Discrimination. *Cognition* 82, B101--B111.

[20] Polka, L., Werker, J. F. (1994). Developmental Changes in Perception of Nonnative Vowel Contrasts. *Journal of Experimental Psychology: Human Perception and Performance* 20(2), 421–435.

[21] R Development Core Team. (2004). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. http://www.r-project.org

[22] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics* 6(2), 461–464. https://doi.org/10.1214/aos/1176344136

[23] Swingley, D. (2009). Contributions of Infant Word Learning to Language Development. *Philosophical Transactions of the Royal Society B Biological Sciences* 364, 3617–3632.

[24] Venables, W. N., Ripley, B. D. (2002). *Modern Applied Statistics with S (Fourth)*. New York: Springer.

[25] Werker, J. F., Tees, R. (1984). Cross-Language Speech Perception: Evidence for Perceptual Reorganization During the First Year of Life. *Infant Behavior and Development* 7, 49–63.