Normalization may be ineffective for phonetic category learning

Kasia Hitczenko[1], Reiko Mazuka[2], Micha Elsner[3], & Naomi Feldman[1,4]
[1]University of Maryland, Linguistics, [2]RIKEN Center for Brain Science,
[3] The Ohio State University, Linguistics, [4] University of Maryland, UMIACS

Sound categories often overlap in their acoustics, which can make phonetic category learning difficult [1]. Several studies have suggested that normalizing acoustics relative to context would improve category separation [2, 3, 4]. However, recent work shows that normalization is ineffective for learning the Japanese vowel length contrast from spontaneous child-directed speech, if categories are not yet known [5]. We analyze why this discrepancy occurs and find that it arises from differences between spontaneous speech and controlled lab speech. Thus, normalization is unlikely to help reduce category overlap in real, naturalistic phonetic learning situations.

Most previous work studied normalization on controlled lab speech, while [5] studied spontaneous speech. We applied the same analyses from [5] to Japanese controlled lab speech from [6] in which mothers teach their infants nonce words. This speech was either read ('Werker Read') or spontaneous ('Werker Spontaneous'), but even the spontaneous speech was less naturalistic than the data used in [5]: the task from [6] controlled how many short/long vowels there were, and what contexts they occurred in. We also replicated the analyses from [5], which were run on data from [7] ('RJMICC Spontaneous'), and compared normalization efficacy on these three datasets.

Following [5], we used logistic regression to separate short/long vowels based on either unnormalized acoustic cues (duration, formants), or cues with all available contextual factors normalized (regressed) out (Table 1). We further computed an upper bound on normalization performance by running a third logistic regression using cues with the best possible subset of contextual factors normalized out (determined by cross-validation on a training set). Contextual factors were broadly construed to include speaker, vowel quality, neighboring sounds, etc.

Normalizing with all factors was ineffective on all three corpora, but the normalization upper bound was much better, relative to unnormalized, on Werker Read and Spontaneous than on RJMICC Spontaneous (Table 2). This means normalization can be helpful on lab-controlled speech or more balanced spontaneous speech, but is relatively ineffective on the naturalistic spontaneous speech that infants learn from. The poor performance on spontaneous speech held even when we replaced linear regression with a neural network, which can learn more complex normalization functions.

We next investigated why normalization didn't work in specific cases, and found that context-specific imbalances between short and long vowels made normalization ineffective. Equation 1 quantifies the extent to which imbalances impede normalization.

$$\left(\mu^{norm}_{l=\text{long}} - \mu^{norm}_{l=\text{short}}\right) - \left(\mu^{unnorm}_{l=\text{long}} - \mu^{unnorm}_{l=\text{short}}\right) = \sum_j \left[\frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}} - \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}}\right]\left[\frac{N_{l=\text{long},c=j}}{N_{c=j}}\mu_{l=\text{long},c=j} + \frac{N_{l=\text{short},c=j}}{N_{c=j}}\mu_{l=\text{short},c=j}\right] \quad (1)$$

In Equation 1, $N_{l,c}$ is the number of vowels of length l in context c and $\mu_{l,c}$ is the mean of vowels of length l in context c. This equation shows how the distance between the overall short and long vowel category means changes after normalization. Normalization will pull the means apart when this value is positive and will push the means together when this value is negative. The equation reveals that all else being equal, if the proportion of long and short vowels differs across contexts, normalization will actually decrease the distance between category means. This occurs because

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 369-370.
New York City, New York, January 3-6, 2019

imbalances in a particular context artificially shift the mean duration of vowels in that context (i.e. having disproportionately many long vowels increases the mean).

Our results call into question the role of normalization in helping learners deal with acoustic variability in category learning. Learners whose input contains phonotactic constraints, phonological alternations, and other context-specific imbalances (which are common in language) would do better to use context as an informative top-down cue to category membership, rather than normalize it out.

Werker Read	Werker Spontaneous	RJMICC Spontaneous
Vowel Quality	**Vowel Quality**	Vowel Quality
Speaker	**Speaker**	**Speaker**
Previous Sound	**Previous Sound**	Previous Sound
Following Sound	Following Sound	Following Sound
Prosodic Position	**Prosodic Position**	**Prosodic Position**
F0	**F0**	Accented
		Duration of Adjacent Sounds
		Part of Speech

Table 1: The full set of contextual factors available for each dataset, with factors that were included in the normalization upper-bound shown in bold.

Data	Model	Accuracy	BIC (Lower is better)
Werker Read	Unnormalized	91.4	246
	Normalized (all factors)	86.1	399
	Normalized (best factors)	95.1	105
Werker Spontaneous	Unnormalized	82.9	1072
	Normalized (all factors)	78.5	1219
	Normalized (best factors)	90.0	869
RJMICC Spontaneous	Unnormalized	91.2	28716
	Normalized (all factors)	91.2	30990
	Normalized (best factors)	91.2	28122
	Normalized (neural network)	91.2	28188

Table 2: Efficacy of normalization, averaged across 10 runs. Unnormalized RJMICC performs well, despite being spontaneous, because 90.9% of the vowels are short.

References

[1] R. A. Bion, K. Miyazawa, H. Kikuchi, and R. Mazuka, "Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech," *PloS ONE*, vol. 8, no. 2, 2013.

[2] B. Dillon, E. Dunbar, and W. Idsardi, "A single stage approach to learning phonological categories: insights from Inuktitut," *Cognitive Science*, vol. 37, no. 2, pp. 344–377, 2013.

[3] J. Cole, G. Linebaugh, C. Munson, and B. McMurray, "Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach," *Journal of Phonetics*, vol. 38, no. 2, pp. 167–184, 2010.

[4] B. McMurray and A. Jongman, "What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations," *Psychological Review*, vol. 118, no. 2, pp. 219–246, 2011.

[5] K. Hitczenko, R. Mazuka, M. Elsner, and N. H. Feldman, "How to use context to disambiguate overlapping categories: The test case of Japanese vowel length," *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, vol. 40, 2018.

[6] J. F. Werker, F. Pons, C. Dietrich, S. Kajikawa, L. Fais, and S. Amano, "Infant-directed speech supports phonetic category learning in English and Japanese," *Cognition*, vol. 103, no. 1, 2007.

[7] R. Mazuka, Y. Igarashi, and K. Nishikawa, "Input for learning Japanese: RIKEN Japanese mother-infant conversation corpus," *The Technical Report of the Proceedings of the Institute of Electronics, Information and Communication Engineers*, vol. 106, no. 165, pp. 11–15, 2006.