The conditioning of surprisal and many-to-one mappings in speech variation: Speaker-orientation and listener-orientation

INTRODUCTION In the literature on speech variation, information-theoretic[1] measures have been found to predict several types of systematic variations (e.g., duration[2], merger[3], lenition[4]). A common finding is that more informative/unpredictable items maintain higher contrasts/are more resilient to variation, supporting listener-oriented perceptual recoverability constraints. However, it is not yet clear how widespread these effects are. First, it remains to be seen whether the effects persist across the board in continuous speech, unconstrained by specific types of systematic variations. Second, since previous studies focused on acoustic data, the effect on articulation, if any, is unknown. While the literature supports intelligibility-based listener-orientation, potential influences of speaker-oriented articulatory efficiency pressures have also been examined[5]. This study entertains the possibility of both forces coexisting. Specifically, we ask: 1) Will information-theoretic (here, surprisal) effects be found across the board in acoustics and articulation, and 2) Could the speech target's flexibility allow for a stronger surprisal effect in acoustics, in consideration of listener-orientation, while decreasing it in articulation, in the face of speaker-oriented pressures for articulatory economy?

METHODS This study examined the contextual surprisal of phone bigrams (phone1-phone2) in a word and its effect on the distinctiveness between the two phones (i.e., phone2's resilience to variation). Importantly, the influence of the target phone's (phone2) flexibility (i.e., attaining a speech target through multiple viable articulatory synergies) on such an effect was also investigated. *Data* We used published articulatory/acoustic data in American/British English and French, collected with ultrasound/optical lip imaging, EMA, and MRI (Table 1). *Distinctiveness* The distinctiveness between two phones was calculated as the mean pairwise Euclidean distance across the temporal dimension between the two phones' articulatory/ acoustic matrices. Acoustic matrices were extracted as the audio's MFCCs. Articulatory matrices were extracted through dimensionality reduction with convolutional autoencoders trained to embed the video into important dimensions through video reconstruction (for imaging data) or as x, y, and/or z coordinates (for EMA/annotated MRI). *Flexibility* A phoneme's flexibility was taken as the mean entropy of the articulatory matrix conditioned on the given acoustic matrix, estimated through mixture density networks[6]. *Surprisal* The contextual surprisal of a bigram was calculated as its negative log probability within the word based on the frequencies in public corpora[7, 8, 9]. *Statistical analysis* We used linear mixed-effects models to predict distinctiveness, with main effects of surprisal, phone1 and phone2 flexibility, distinctiveness type (articulatory vs. acoustic), and phone1 and phone2 frequency, and random intercepts and slopes of surprisal and phone2 flexibility by word and subject.

RESULTS We found significant positive surprisal effects on distinctiveness across all datasets, except for FR-MRI, in both articulation and acoustics. Such surprisal effects were stronger in acoustics in AE-MRI and AE-MRI-annotated, but stronger in articulation in BE-US and AE-EMA (Figure 1). Importantly, phoneme flexibility, through interaction with surprisal and modality (articulation vs. acoustics), enhanced the surprisal effect in acoustics and decreased it in articulation across all five datasets (Figure 2).

DISCUSSION Our results further support listener-oriented intelligibility-based information-theoretic conditioning of speech variation, and show that such conditioning is discoverable in a bottom-up, unsupervised manner, unconstrained by specific variation types. We find a unified pattern in the differences in such conditioning between articulation and acoustics by considering target flexibility: when the target phoneme is more flexible, surprisal effects were further enhanced in acoustics, but decreased in articulation. Thus, while listener-orientation is dominant in such conditioning, the flexibility of the target phone serves as a window for speaker-orientation to compete with listener-oriented pressures.

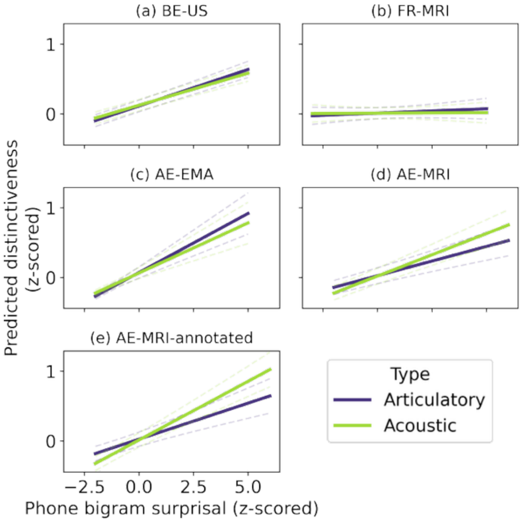

Figure 1. Surprisal effects on distinctiveness (dotted lines indicate 95% CI).

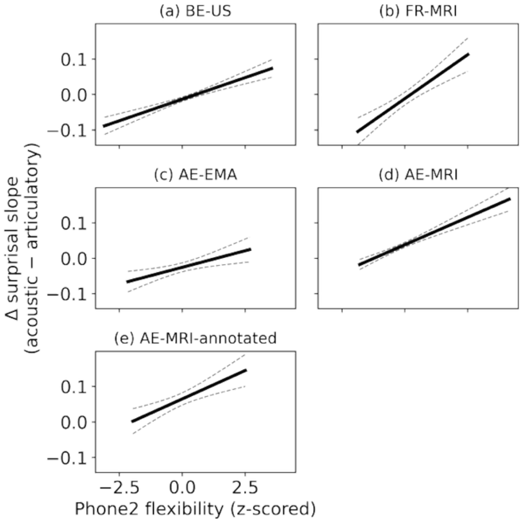

Figure 2. Difference of surprisal effects between articulation and acoustics under the effect of flexibility (dotted lines indicate 95% CI).

Table 1. Summary of the datasets used in this study.

	Dataset	Language	Instrument	# participants	# bigram
BE-US	Tongue and Lips Corpus[10]	British English	Ultrasound/Optical lip imaging	41	251,722
FR-MRI	ArtSpeechMRIfr[11]	French	MRI	10	42,644
AE-EMA	USC-TIMIT[12] & 75-Speaker Annot-16[13]	American English	EMA	4 (subset)	81,698
AE-MRI			MRI	17	232,088
AE-MRI-annotated				7 (subset)	42,218

Table 2. Summary of the three key measurements in this study.

	Type	Measured as	Interpretation
Distinctiveness	acoustic	Mean Euclidean distance across the frames in MFCCs	Amount of acoustic/articulatory difference between the phone1 and phone2 of the bigram (effectively, phone2's resilience to variation as a function of phone1)
	articulatory	Mean Euclidean distance across the pairwise comparison of the frames in raw coordinates/model-extracted matrices	
Contextual predictability		Contextual surprisal	Predictability of the phoneme bigram given the preceding phonemic context in the word
Flexibility		Conditional entropy of the articulatory matrix given the acoustic matrix	Degree of freedom between the viable articulatory synergies and the speech target

References

[1] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27* , 379–423.

[2] Baker, R. E., & Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech, 52*(4), 391–413. doi: 10.1177/0023830909336575

[3] Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition, 128*, 179–186.

[4] Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language, 93*(3), 569–597.

[5] Xu, Y. , & Prom-On, S. (2019) . Economy of effort or maximum rate of information? Exploring basic principles of articulatory dynamics. *Frontiers in Psychology, 1*0, 2469.

[6] Zen, H., & Senior, A. (2014). *Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis.* In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

[7] BNC Consortium. (2007). T*he British National Corpus, version 3* (BNC XML Edition). Oxford: Oxford University Computing Services.

[8] Langlais, P.-C., de Courson, B., & Azoulay, B. (2024). French Public Domain Books (Dataset) [Dataset]. Hugging Face.

[9] Open American National Corpus. (2008). Open American National Corpus (OANC) [Dataset]. American National Corpus Project.

[10] Ribeiro, V., & Laprie, Y. (2022, September). *Autoencoder-based tongue shape estimation during continuous speech.* In *Proceedings of Interspeech 2022* (pp. 86–90). ISCA.

[11] Douros, I.K., Felblinger, J., Frahm, J., Isaieva, K., Joseph, A.A., Laprie, Y., Odille, F., Tsukanova, A., Voit, D., Vuissoz, P.-A. (2019) A Multimodal Real-Time MRI Articulatory Corpus of French for Speech Research. *Proc. Interspeech 2019*, 1556-1560.

[12] Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.-C., Zhu, Y., Goldstein, L., Byrd, D., Bresch, E., Ghosh, P., Katsamanis, A., & Proctor, M. (2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *The Journal of the Acoustical Society of America, 136*(3), 1307–1311.

[13] Shi, X., Zhang, Y., Lu, Y., Ma, M., Feng, T., Toutios, A., Hsu, H., Goldstein, L., Narayanan, S. (2025) 75-Speaker Annot-16: A benchmark dataset for speech articulatory rt-MRI annotation with articulator contours and phonetic alignment. *Proc. Interspeech 2025*, 2175-2179.