Speakers matter more than listeners in sound identification accuracy

Daniel Aalto[1], Naomi Gurevich[2]

[1]Communication Sciences and Disorders, University of Alberta [2]Communication Sciences and Disorders, Purdue University Fort Wayne

PURPOSE

Theoretical context:

Communication relies on language: a hierarchical system operating on discrete units (e.g., phonemes) transmitted through speech production and perception systems[1]. While many factors influence the decoding accuracy, their relative contributions are not clear.

Speakers articulate speech sounds along a continuum of clarity (H&H theory) modulated by communicative constraints[2], e.g., syllable final sounds are more reduced and less accurately identified than syllable initial[3,4]. *Listeners* identify sounds using both acoustic **(bottom-up)** and linguistic **(top-down)** information. Sounds vary in their frequency in lexicon (/t/ is more common than /h/) and their acoustic salience.

Research question:

What are the relative contributions of listeners, speakers, and phonology to the variability in sound identification accuracy?

METHODS

Speech materials:
- 21 native English speakers read 308 phonetically rich words[5]
- 384 triphones were extracted resulting in 8,064 audio clips

Listening task:
- 22 native English listeners each transcribed ≈745 audio clips based on a random Latin square design resulting in 16,408 transcriptions

Accuracy analysis:
- transcriptions were manually compared to the target sounds

Statistical analysis:
- mixed effects logistic regression (lme4 in R)
- fixed effects: sound frequency (5 levels); context prominence; stress, number of phonemes
- random effects: speaker, listener, triphone

RESULTS

Accuracy Fixed effects	Level	Odds ratio	Estimate	Std. Error	z value	p value
Intercept			4.66	0.32	14.6	2.0E-16
Sound frequency (ref: very high /r,t,n,s,l,k/)	high /p,b,m/	0.44	-0.82	0.20	-4.0	5.7E-05
	medium /f,b,ʃ, v/	0.46	-0.79	0.19	-4.1	4.9E-05
	low /g, w, z, ʤ, ŋ, j, ʧ, h/	0.30	-1.22	0.18	-6.8	1.1E-11
	very low /θ, ð, ʒ/	0.13	-2.03	0.26	-7.7	1.2E-14
Context prominence (ref: #_V, V_V, C_V)	#_C	0.96	-0.05	0.26	-0.2	0.86
	V_C,V _#, C_C, C_#	0.45	-0.81	0.14	-5.7	1.3E-08
Stress (ref: stressed, polysyllabic word)	(stressed) monosyllabic	1.00	0.00	0.20	0.0	0.98
	unstresed polysyllabic	0.56	-0.59	0.16	-3.7	0.00018
Number of phonemes		0.87	-0.14	0.04	-3.3	0.00083

Random effects:
- σ(listeners) = 0.15 (10%)
- σ(speakers) = 0.28 (18%)
- σ(triphones) = 1.12 (72%)

Interpretation:
- Listener variability is smaller than speaker variability
- Variability of the triphones was largest although none of the fixed effects were related to the participants
- σ(X) could be interpreted as a likelihood for a communication breakdown to be due to X

Top-down/predictability effects
- More frequent sounds are identified more accurately
- high and medium frequency groups equally well identified

Bottom-up/articulatory effects
- Speech sounds in shorter words, stressed syllables, or prominent positions (pre-vocalic/word initial) are identified more accurately
- Higher identification accuracy could be due to slower/clearer speech

DISCUSSION

Linguistic factors:
Sound frequency[6], context prominence[3], stress[7] effects aligned with previous literature

Triple inference:
Statistically robust and feasible with only 43 participants to disentangle language, speaker, and listener effects

Ecological validity:
Multiple speakers make the task difficult enough so that noise is not needed

Future directions:
What articulatory factors (e.g., rate, amount of reduction) contribute to identification/intelligibility?

Clinical significance:
The results support a hierarchical approach to Intelligibility interventions[8]

REFERENCES

[1]Keyser, S. J., & Stevens, K. N. (2006). Enhancement and overlap in the speech chain. Language, 82(1), 33-63.
[2]Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Speech production and speech modelling (pp. 403-439). Dordrecht: Springer Netherlands.
[3]Recasens, D. (2004). The effect of syllable position on consonant reduction (evidence from Catalan consonant clusters). Journal of Phonetics, 32(3), 435-453.
[4]Woods, D., Yund, E. W., Herron, T. J., & Cruadhlaoich, M. A. (2010). Consonant identification in consonant-vowel-consonant syllables in speech-spectrum noise. The Journal of the Acoustical Society of America, 127(3), 1609-1623.
[5]Gurevich, N., & Kim, H. (2023). Development of novel speech stimuli with phonetic coverage and phonemic balance. *Perspectives of the ASHA Special Interest Groups, 8*(2), 424-437.
[6]Moates, D. R., Watkins, N. E., Bond, Z. S., & Stockmal, V. (2006). Frequency effects in phoneme processing. The Journal of the Acoustical Society of America, 120(5), 3252-3252.
[7]Cilibrasi, L., Stojanovik, V., & Riddell, P. (2015). Word position and stress effects in consonant cluster perception and production. Dyslexia, 21(1), 50-59.
[8]Gurevich, N., & Kim, H. (2022). Phonetics and Phonology: The phonetics and phonology of intelligibility: The functional importance to intelligibility of speech sounds. In Clinical Applications of Linguistics to Speech-Language Pathology (pp. 111-125). Routledge.

CLAIM: LANGUAGE IS THE MOST VULNERABLE COMPONENT IN THE SPEECH CHAIN!