CellPress

The social weight of spoken words

Meghan Sumner

Department of Linguistics, Margaret Jacks Hall, Bldg 460, Stanford University, Stanford, CA 94305-2150, USA

Speech serves a linguistic function, cueing sounds and words, and a social function, cueing talkers and their social attributes. Listeners readily map sound patterns in speech to social representations. This mapping introduces social biases on the recognition and encoding of sound patterns produced by different groups and individuals.

Speech is highly variable. A spoken word is never uttered the same way twice. This variation in speech is not random. It is highly predictable, both in terms of its linguistic and its social structure. Variation provides information about not only sounds and words, but also a talker's age, gender, accent, emotion, and style. Phonetic variation, then, is a multifaceted information source. As humans, we communicate information about ourselves through variable acts every day. We do this by varying speed, posture, and style in our gait [1] and by varying our use of language [2]. Both are dynamic processes. Just as humans automatically map co-present cues in a gait to recognize social attributes of walkers, we might expect that humans automatically map co-present cues in speech to recognize social attributes of talkers.

However, the recognition of these attributes depends greatly on how we encode speech events in the first place. When hearing spoken language, we, as listeners, experience a constant stream of speech along with co-present sensory information and expectations. Consider career day. A theoretical physicist enters the room to talk about his road to academia and his research. His gestures provide examples of how a physicist acts. His words and, importantly, the way he utters those words, provide examples of how a physicist talks and sounds. How each student internalizes this event will greatly depend on their own goals, desires, and social experiences. These, in turn, will affect the ways in which this event is encoded. A student who has wanted to be a physicist her entire life, who has been encouraged to do so, and who identifies with the talker will listen differently compared with a student who sees this profession as a profession for men, who has been encouraged to pursue her love of books, and who sees no social connection to the talker. The former will have all cognitive faculties tuned toward the event, absorbing patterns; likely remembering some key sentences or phrases exactly. The latter might be calculating the remaining time, looking out the window, doodling; all while the talker's words continue to stream to her. In this case, she is likely to remember the gist of what was said, but not the exact phrasing.

Corresponding author: Sumner, M. (sumner@stanford.edu).

Keywords: speech perception; spoken word recognition; phonetic variation; social variation.

1364-6613/

© 2015 Elsevier Ltd. All rights reserved. http://dx.doi.org/10.1016/j.tics.2015.03.007

Talker information conveyed through speech is one of many contextualizing elements that guide behavior. In an early study showing discriminatory behavior cued by talker variation, a single speaker produced different voice guises in response to rental advertisements. These guises were based on Standard American English, Chicano American English, and African American English speech patterns [3]. Callbacks from lessors were significantly higher for the Standard American English guise than for the others. Therefore, listeners use information conveyed solely by a voice as one piece of information guiding their decision to whom to rent the apartment.

Most people would not be surprised to learn that listeners discriminate based on such social cues, but they might assume that these biases arise late during the processing of spoken language. It appears that voice cues activate social representations fast and early during the process of spoken language understanding. One consequence of this early activation is that it provides an outlet through which our social biases may modulate the allocation of cognitive resources, influencing the encoding and retention of auditory information.

Much work in speech perception and spoken word recognition over the past 30 years has centered on the linguistic function of speech, investigating how listeners map a variable speech signal to (seemingly) discrete linguistics categories such as sounds and words. This research has converged on the finding that variation is critical to the composition of our representations, and facilitates language processing (e.g., [4]). However, phonetic variation in speech conveys information beyond the literal translation of speech to sounds and words. Speech exhibits patterned, conventionalized variation for words, sounds, and their acoustic features, used by speakers to create style and identity [5]. In this way, phonetic variation is a two-sided coin, serving a social function in language, as well as a linguistic one. Yet, our understanding of the effects of this social function of phonetic variation on the perception, recognition, and understanding of spoken words is limited compared with what we know about the mapping of speech to linguistic categories and about how speakers use language in production to create social meaning.

For evidence that social information conveyed through a talker's voice triggers social representations fast and early, we turn to behaviors beyond the literal translation of speech to sounds and words. Words spoken by women with typical female voices are identified more quickly than those with atypical female voices [6]. Happy words are recognized more quickly when uttered with happy versus sad prosody [7]. In addition, words typically uttered by older adults are recognized more quickly when the speaker is an older adult rather than a younger adult [8]. These

CrossMark

studies tie talker voice to our word-level experiences in an exemplar lexicon, where clusters of word forms are activated via an acoustic similarity metric [6].

Yet, talker information is processed even before we access our lexicon [7]. Children can learn to associate talker voice with color preferences [9] and event-related brain potentials (ERP) measurements show that adult listeners process the fundamental frequency (one major acoustic cue to talker gender perception) of syllables both for its linguistic and social functions [10].

One consequence of early access to social representations is the potential for our biases to influence how we attend to speech in the first place. Just as the color of one's skin or education level might, however unjustly, influence our behavior, voice cues activate the same social biases and influence behavior in a similar manner. This process is referred to as social weighting [11].

We can think of social weighting as a feedback loop where social representations modulate the allocation of cognitive resources to speech processing (from voice or other cues; as in the classroom example). In other words, all experiences with spoken language are not treated equally. Consider an example where attention is drawn to the specific word form. A mother might slow down and produce a rare [t] in the word *center* to aid a child in spelling (contrary to intuition, words such as these typically sound like *sen-ner*). The child does not explicitly choose to process the form deeply. Rather, deep encoding results from a social situation that encourages form processing. Over the years, effects of orthography, metalinguistic commentary, and other experiences compound, yielding a perceived standard form that is not reflected in global rates of production. This results in strongly encoded forms that, despite being rarely experienced, yield a robust form-based lexical representation. This representation may be equally robust as one derived from more frequently encountered, but weakly encoded, forms of words (such as *center* without [t]).

Although the frequency with which we experience linguistic units is influential in models of spoken language processing, social weighting explains how things we rarely hear are understood as well as things we typically hear. This recognition equivalence occurs within different registers of a single accent [12] and across speakers with different accents [13]. In contrast to equivalence between a typical and rare form (as the two instances of *center* above), we might also expect differences in processing equally rare forms produced by talkers from different social backgrounds [14]. If quantitative exposure to speech variation alone predicted how well we understand spoken words, a cost should be associated with all speech produced by talkers with accents different from one's own. Although costs have been associated with out-of-accent talkers with stigmatized regional [14] or foreign accents [15], analogous costs are not found when listeners hear an out-of-accent, but prestigious talker. In this case, listeners recognize words produced by an out-of-accent talker on par with those produced by a within-accent talker.

The asymmetries do not stop there. Moving from immediate processing to memory for spoken words, perceived standard forms (e.g., *sen-ter*) are remembered better than typical, nonidealized forms (e.g., *sen-ner*) or atypical, nonstandard forms [12]. Different populations of listeners remember socially idealized productions of words (e.g., General American English *slend-er*) better than forms associated with a stigmatized dialect (e.g., New York City, *slend-uh*) [13]. Remarkably, this memory difference is found even among listeners who themselves speak the stigmatized dialect. Finally, when recalling spoken words produced by talkers with accents different from the listening population, listeners remember the gist of what was said for a talker with a stigmatized accent (e.g., New York City) and exactly what was said for a talker with a prestigious accent (e.g., Southern Standard British English) [14].

As pattern recognizers, we learn to map acoustic bundles to social representations through experience with copresent auditory and visual cues in the world around us. One consequence (of many) of this mapping is that our biases influence the way we listen to speech. The fast mapping of speech to social categories may be helpful in navigating the complex process of understanding spoken words. However, it also introduces biases and discriminatory behavior early in the spoken word recognition process, much earlier than we might hope.

Acknowledgments

I am grateful to Annette D'Onofrio, Seung Kyung Kim, and Kevin McGowan for valuable comments and feedback. This material is based in part upon work supported by the National Science Foundation under Grant Numbers 0720054 and 1226963 made to M.S. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

1 Cutting, J.E. and Proffitt, D.R. (1981) Gait perception as an example of how we may perceive events. In *Intersensory Perception and Sensory Integration* (Walk, R. and Pick, H., eds), pp. 249–273, Plenum
2 Eckert, P. (2012) Three waves of variation study: the emergence of meaning in the study of variation. *Annu. Rev. Anthropol.* 41, 87–100
3 Purnell, T.C. *et al.* (1999) Perceptual and phonetic experiments on American English dialect identification. *J. Lang. Soc. Psychol.* 18, 10–30
4 Beddor, P.S. *et al.* (2013) The perceptual time course of coarticulation. *J. Acoust. Soc. Am.* 133, 2350–2366
5 Irvine, J. (2001) Style as distinctiveness: the culture and ideology of linguistic differentiation. In *Stylistic Variation in Language* (Eckert, P. and Rickford, J., eds), pp. 21–43, Cambridge University Press
6 Johnson, K. (2006) Resonance in an exemplar-based lexicon: the emergence of social identity and phonology. *J. Phonet.* 34, 485–499
7 Nygaard, L.C. and Queen, J.S. (2008) Communicating emotion: linking affective prosody and word meaning. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1017–1030
8 Walker, A. and Hay, J. (2011) Congruence between 'word age' and 'voice age' facilitates lexical access. *Lab. Phonol.* 2, 219–237
9 Creel, S.C. (2012) Preschoolers' use of talker information in on-line comprehension. *Child Dev.* 83, 2042–2056
10 Kaganovich, N. *et al.* (2006) Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Res.* 1114, 161–172
11 Sumner, M. *et al.* (2014) The socially-weighted encoding of spoken words: a dual-route approach to speech perception. *Front. Psychol.* 4, 1–13
12 Sumner, M. and Samuel, A.G. (2005) Perception and representation of regular variation: the case of final-/t/. *J. Mem. Lang.* 52, 322–338
13 Sumner, M. and Samuel, A.G. (2009) The role of experience in the processing of cross-dialectal variation. *J. Mem. Lang.* 60, 487–501
14 Sumner, M. and Kataoka, R. (2013) Effects of phonetically-cued talker variation on semantic-encoding. *J. Acoust. Soc. Am.* 134, EL485–EL491
15 Lev-Ari, S. and Keysar, B. (2010) Why don't we believe non-native speakers? The influence of accent on credibility. *J. Exp. Soc. Psychol.* 46, 1093–1096