Language, Cognition and Neuroscience, 2015
Vol. 30, No. 3, 273–286, http://dx.doi.org/10.1080/23273798.2014.906636

Routledge
Taylor & Francis Group

Analogy and disanalogy in production and perception of speech

Robert E. Remez*

Department of Psychology and Program in Neuroscience & Behavior, Barnard College, Columbia University, New York, NY, USA

A varied psychological vocabulary now describes the cognitive and social conditions of language production, the ultimate result of which is the mechanical action of vocal musculature in spoken expression. Following the logic of the speech chain, descriptions of production have often exhibited a clear analogy to accounts of perception. This reciprocality is especially evident in explanations that rely on reafference to control production, on articulation to inform perception, and on strict parity between produced and perceived form to provide invariance in the relation between abstract linguistic objects and observed expression. However, a causal account of production and perception cannot derive solely from this hopeful analogy. Despite sharing of abstract linguistic representations, the control functions in production and perception as well as the constraints on their use stand in fundamental disanalogy. This is readily seen in the different adaptive challenges to production – to speak in a single voice – and perception – to resolve familiar linguistic properties in any voice. This acknowledgement sets descriptive and theoretical challenges that break the symmetry of production and perception. As a consequence, this recognition dislodges an old impasse between the psychoacoustic and motoric accounts in the regulation of production and perception.

Keywords: speech perception; speech production; language system architecture

At some moments in your busy life, the stream of awareness of yourself, of objects, of your companions, of events, of the receding past and of the looming future is likely to include a spur to action… and, you say something. Issuing a message differs in kind from other acts you might commit, like bringing groceries home or climbing Mount Everest. Although a linguistic expression takes physical form through vocalisation, the aim of such acts is to represent your intentions to another, and the physical acts appear merely to be the means to make your intentions public. In our era, we are often admonished to see the conceptual part as the heart of language, the formal devices by which an individual's communicative intentions are representable at all. Yet, grammatical representations must be expressible or there is no motive to compose such linguistic formulae in the first place. This is not a subtle point. It says that a talker with semantic intentions must know how to compose a communicable linguistic form, and how to give voice to it; and, a listener must be able to notice the useful properties of vocalisation and know how to resolve its linguistic components. From this perspective, there can be no language without expression, which poses a scientific challenge to understand this amalgam of conceptual and expressive resources. Neither portion has proven to be the easier one to explain.

With every talker a listener, the functions of production and perception stand in reciprocal relation, cognitively. Historically, accounts of this relation have ranged widely. At one ideological extreme, frankly reductionist approaches have viewed one side of the reciprocality deriving from the other. Despite this explanatory opportunity, some accounts have seen the convergence of production and perception as a consequence of sensory and motoric resources applied in parallel to speech. Is there a winning view to be found among these contending claims? By acknowledging that production and perception serve a matching linguistic aim, it is reasonable to anticipate analogies between these cognitive functions. Yet, recognising the differing inherent roles of production and perception permits us to anticipate differences in these functions, and to admit that disanalogy must likewise be typical of their relation. Here, a retrospective prelude of classic approaches introduces a selective discussion of studies that seem, for now, to be pivotal in understanding the relation between production and perception of spoken language.

A retrospective prelude

Various notions of the relation between production and perception are found in the origins of cognitive psychology. Our common ancestor, Karl Lashley, took the high road, noting that the formal linguistic challenge to speak an ordered series of constituents hardly differed from the perceptual requirement to resolve constituents spoken in order (Lashley, 1951). The mechanistic key to his conceptualization was a spatial neural array of simultaneously

*Email: remez@columbia.edu

available expressive elements, and a temporal array that scanned it, sequentially and metrically, issuing a series of commands to a vocal apparatus that spoke the utterance. This was reciprocal to the perceptual accretion of sequentially heard elements, projecting each into a spatial neural array wherein the relations among simultaneous elements could become apparent. Exchange errors and errors of anticipation proved that his conceptualization of the productive side was astute; delayed binding of form to meaning, contingent on the context established by a series of expressed elements represented simultaneously, proved likewise for perception.

Attention to linguistic form is also characteristic of the approach of Roman Jakobson and Morris Halle whose description of the contrastive nature of a phonological system minimised the unique attributes of production and perception. They wrote:

> To find out what motor, acoustic and perceptual elements of sounds are utilized in a given language, we must be guided by its coding rules: an efficacious physiological, physical and psychological analysis of speech sounds presupposes their linguistic interpretation (p. 33)…. The specification of distinctive oppositions may be made with respect to any stage of the speech event, from articulation to perception and decoding, on the sole condition that the invariants of any antecedent stage be selected and correlated in terms of the subsequent stages, given the evident fact that we speak to be heard and need to be heard in order to be understood… [though the purported] closer relationship between perception and articulation than between perception and its immediate stimulus finds no corroboration in experience (p. 34)… we are not concerned with substituting an acoustic classification for an articulatory one but solely in uncovering the most productive criteria of division for both aspects (p. 36). (Jakobson & Halle, 1956)

In short, the challenge to maintain a system of coherent phonological contrasts whether binary or *n*-ary depends on reliable production and recognition alike, and requires each modality to conserve linguistic distinctions.

Or, so it had seemed until instrumental means were used to calibrate the acts of production and their acoustic consequences. The goal of identifying invariant relations among the links in the speech chain remained hypothetical yet out of reach, and the search gave rise to prominent reductionist accounts. In the Motor Theory of speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), departures from strict invariance in the relation among phoneme, articulation and acoustic pattern were attributed to coarticulation of phonemes. That is, although a phoneme may be conceptualised as a discrete linguistic object, available for sequencing and clustering with others according to phonological principles – not to mention the erroneous commutivity and perseveration described by Lashley – kinematic and acoustic measures revealed that its production is anything but discrete. Instead, each phoneme is produced with a time-course, and because the motor expression of a phoneme is imbricated temporally with preceding and following phonemes, no single phoneme is produced free of the influence of its sequence. According to Motor Theory, specific vocal acts and the acoustic effects they produce are a code exhibiting neither linearity nor invariant correspondence to phonemes; as a result, the inverse function required for the perception of phonemes from acoustics must invoke understanding of coarticulation, hence, perception presumes an understanding of production. Although the Motor Theory of speech perception exists in several versions (Liberman & Mattingly, 1985), all retain this emphasis: the segmental origin of the continuously evolving acoustic signal of speech can be recovered only by incorporating the dynamics of articulation, to render the phonetic series free of the blending effects of coarticulation of phonemes in production. The inclusion of special knowledge of articulatory dynamics led to the description of speech perception as a biologically specialised function (Mattingly & Liberman, 1988), making speech perception an epitome of cognitive modularity (Fodor, 1983).

The antipode of a special account is a general account, and there have been many attempts to explain the phenomena of production and perception of speech intrinsically, without referring production to the perceptible, nor perception to the producible. Regarding production, there are significant problems for an account positing perceptual regulation of motor performance at fine temporal grain. If, overall, the aim of articulation is to produce an acoustic signal composed of resolvable linguistic properties, the procession of individual articulatory events outstrips the perceptual ability to monitor production while it occurs, as Lashley (1951) first noted. For this reason, speech has appeared to be produced with open-loop control, barely disrupted by noise that renders speech inaudible (Lane, Catania, & Stevens, 1961), or by lidocaine that abolishes orofacial taction (Borden, Harris, & Oliver, 1973), or by nerve blockade of intrafusal muscle fibres that eliminates muscle sense (Abbs, 1973). It needs no mention that the blind speak fluently and articulately. Production might be inspired or even motivated perceptually, but is not regulated closely by reafference.

In complementary fashion, perceptual explanations of speech recognition have appealed to a common stock of cognitive functions. Indeed, identifying the phonemic type of an acoustic token is viewed in this perspective as an act of auditory categorization, whether the method invokes schemas, prototypes, or recognition by components, with or without conscientious application of statistical polish. Among these is an account by Massaro (1994) in which acoustic signal elements are associated conditionally with specific phonemes, and perceptual functions derive the likeliest phoneme at any instant from their differential base rates and the immediate signal conditions. This

approach has also been combined with peripheral auditory modelling to estimate a likely preliminary representation of speech, and information theory to model the uncertainty attributable to exposure norms and intrinsic differences in distinctiveness. Among such accounts it is rare to encounter premises pertaining to the causes of speech sounds in articulatory acts. Instead, the focus rests on the categorization of the acoustic samples available to a listener, whether these are sought as superficial and elemental, as in historical discussions of the speech cues, or described as more abstract patterns, as in higher-order auditory categories.

An evolutionary vignette about perceptual multistability

Before concluding that the relation between production and perception is accountable by one or another of these explanations, it will be useful to review a deferred albeit basic concern. Specifically, it would be beneficial to identify the kind of relation that ties linguistic properties to their physical expression (Lane, 1968). After all, the prospect of reduction of any kind depends on the truth of the claim that phonemes are really just gestures of vocalisation, or that phonemes are really just categories of sounds. In this respect, an analogy between phonemes and integers might be helpful, and to give the story some contemporary spice, consider it in an evolutionary setting.

In this version, it was a dark and stormy night long ago when a group of our ancestors (see Figure 1A) sat around a Palaeolithic fire examining some litter composed of animal bones (see Figure 1B). No two are alike; one bone was longer, one shorter, one lighter, one heavier, one straighter, one more bowed, one broken at the end and one shattered in the shaft – you get the idea. By inspecting the physical properties of the bones, the salience of their exquisitely graded variation in form is promoted. This is an obvious and direct way to appreciate the material properties of objects and events and makes use of one stable organisation of attention. Yet, by ignoring the physical properties that distinguish a femur from a tibia, a bone could become a marker equivalent to any other, a counter standing for an integer despite the material attributes distinguishing each one during the prior inspection. This form of attention pivots on an alternate organisation, making use of the inherent multistability of perceptual analysis.

There is no way to tell from reviewing a group of bones whether attention to their graded variation is obliged, or if relief from that physical focus is warranted in order to promote their interchangeable use as surrogates for a numerical abstraction. An act of mind is required (Quine, 1968), and an intended purpose. Much the same is true of the articulatory and auditory correlates of production and perception, although it has been a long time since a behaviourist claimed that a word or phrase was just a

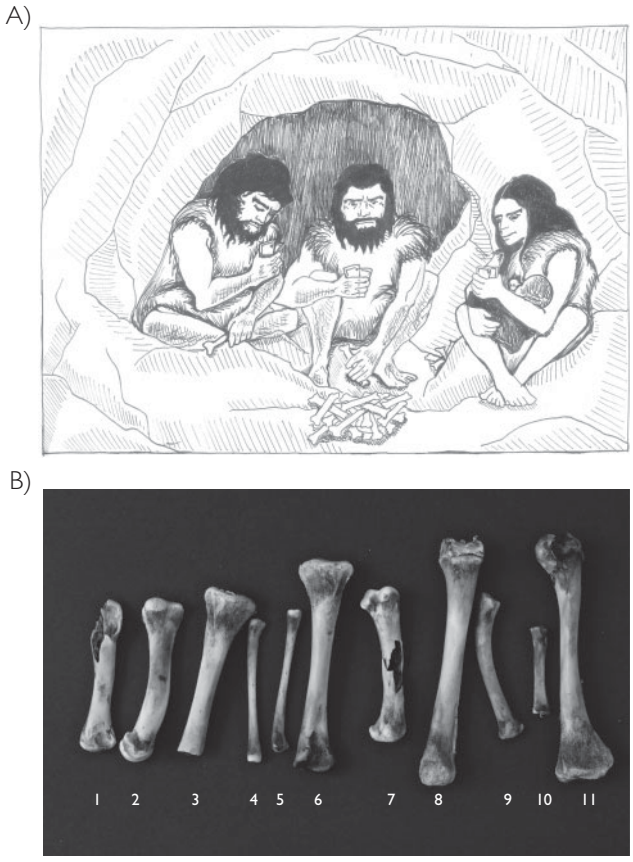

Figure 1. (A) Palaeolithic ancestors depicted in a hypothetical card game, indicating their wagers with opportunistically chosen chips. (B) A collection of several kinds of bones exhibiting graded variation within kind. Despite the uniqueness of each bone, when cavemen played jacks or poker each bone could be used as an equivalent token, on an analogy to the class of integers. See the text for an explanatory vignette.

physical object, spoken or heard. The phonemic contrasts that distinguish the communicable form of words are irreducibly symbolic, and neither more physical nor less conceptual than integers.

From this perspective, a search for physical attributes of articulation or acoustics with the same distribution as each contrast must be fundamentally mistaken. The physical heterogeneity of the set of markers blocks this when bones are used as counters – as in a game of poker – and in speech. This is inherent in the status of phonemes as linguistic entities. In the same way that integers and words are uncommitted to a specific physical form of expression, so, too, are consonant and vowel contrasts uncommitted to the physical form that they can take in production and in perception. To complete the analogy, scientific inspection of articulatory acts, acoustic spectra and states of the auditory system reveals the exquisite variation that accompanies each instance of a phoneme contrast. Moreover, there is no way to tell from the physical form of phonemic expression which properties

are symbolic and which stem from other causes. Under these conditions, one simple description of the logic of the phoneme code is that the production and perception of speech occur as if the commitment to the motor and sensory expression is flexible. Because phonemic contrasts are expressed with very nearly unbounded variation in articulatory, acoustic and auditory form, these phenomena perpetually elude the appeal to a normative rationale which sadly serves as the theory of first and last resort in psychology and behavioural neuroscience.

Points of analogy

To characterise the relationship between production and perception of speech, the convergence of these two modalities makes the speech chain possible, linking talker and listener in form and function. In order for this to occur, there must be basic parity between a linguistic form issued by a talker and recognised by a listener, no matter how the articulatory acts, airborne hazards and auditory samples have mediated the linguistic properties. Technical investigations show that production and perception stand in striking analogy in these ways:

(1) perception and production are effortful;
(2) perception and production converge symbolically;
(3) linguistically effective physical tokens in production and perception vary hugely;
(4) along the dimension *specificity–versatility* there does not seem to be a strong requirement for specificity.

These are considered in turn.

Effortful

The production of speech is voluntary, requiring an intention to act. An utterance indicates sentience, of course, but because conversation cannot be elicited by reflex, the production of speech manifests a talker's deliberateness as well as a focus of cognitive resources on speaking. Although speech is demonstrably replete with errors, some corrected and many simply ignored or missed, the production of speech represents skill established through practice. Its neuromotor organisation is distinct from chewing, deglutition, respiration and other functions supported by the same anatomy and physiology. The brief articulate production observed, however rarely, in coma or sleep is exceptional, and reflects the coordinative potential of incompletely available neural resources. Certainly, the functional character of such extraordinary conditions differs from the ordinary circumstance of deliberate production.

Evidence also supports this designation of deliberateness in the perception of speech, namely, that it requires intention as well as the devotion of cognitive resources. Some studies have pressed this question directly by using sine-wave replicas of speech (Remez, Rubin, Pisoni, & Carrell, 1981). Because sine-wave patterns lack the harmonic structure, pulsing and broadband resonances of speech, they have the quality of simultaneously varying whistles, and listeners rarely group the asynchronously changing pitches as issuing from a single source. Without a hint that the tones compose a kind of synthetic speech, a listener does not notice consonants, vowels or words, instead perceiving a sine-wave utterance as a collection of unrelated tones. An intention to listen for linguistic attributes must accompany the rapidly fading auditory trace. Ordinarily, the vocal quality of natural speech supplies a crucial hint that the wave at the ear is spoken, and supplies it quickly enough to draw resources appropriate for finding and following the modulations of the sound that carry the message. Or, the perceiver might simply see a companion speaking in order to form a belief that sounds at the ear include speech.

Most generally, the difference between listening and passive hearing is significant. Although some recent studies (for instance, Zevin, Yang, Skipper, & McCandliss, 2010) have dubiously presumed an equivalence of deliberate attention and passive exposure to speech, this premise is inconsistent with findings that reveal indiscriminate perceptual organisation – streaming, or grouping – without attention (for example, Carlyon, Cusack, Foxton, & Robertson, 2001). Speech perception requires listening; the mere transduction of speech samples by an auditory system is insufficient (see Remez & Thomas, 2013).

Symbolic convergence

Within a language community, a property that counts as a sign for the talker must also count as one for the listener or communication of the message fails. This symmetry was described as *parity* (Liberman & Mattingly, 1985). Because linguistic constituents are nested, with clauses composed of phrases, phrases composed of words, words composed of syllables and syllables composed of phonemes, it is possible for production and perception of speech to converge at the superordinate level while diverging in detail. If I say /pɨt^{h}ɛiɟo^{w}/ and you say /pɨt^{h}ɑːɟo^{w}/ lexical parity is observed despite the forgiveness required by the segmental lapse. Although the indifference to such departures from isomorphism is itself an intriguing and potentially explosive concern, the success criterion described in the *Speech Chain* (Denes & Pinson, 1963) endures (see Figure 2); the linguistic properties critical in production match those in perception.

Vast variation

The opportunity to observe speech production in progress requires a method to measure the fleet gymnastics of the tongue. The lips and jaw are readily observed, and even the velum presents only a relatively moderate challenge to

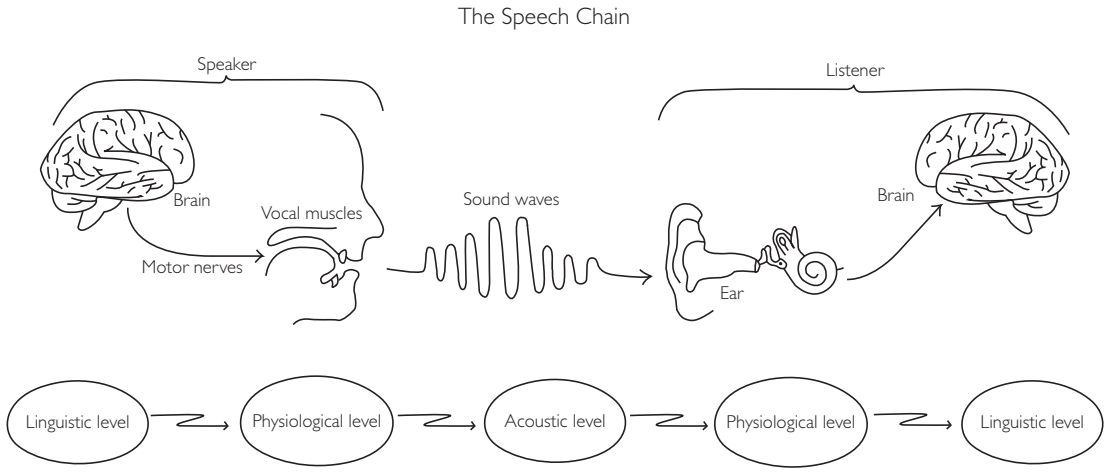

Figure 2. The Speech Chain, redrawn and redacted from the original in Denes and Pinson (1963). Talker and listener match at the linguistic level when spoken communication succeeds, although the chain is mediated by articulatory expression, airborne propagation of sound and auditory sampling.

study. Early radiological methods used to visualise the shapes, placement and motions of the tongue disabused researchers of the hypothesis that a simple relation exists between phoneme and articulation (Moll, 1960; Subtelny, Pruzansky & Subtelny, 1957). Nonetheless, there remains little agreement about the control parameters used in speech production, whether these are positions and motions of vocal articulators or their parts, cavity shapes created in the vocal tract by articulators, or patterns of neural activity arrayed across motor units. One of the longstanding puzzles is the source of the variety of anatomical configurations that realise a single phoneme. In the data-set presented long ago by Öhman (1967), tongue shapes were measured from x-ray motion pictures. The aim was to compare an idealisation of /d/ with the measurements of actual tongue shape and placement. Even ignoring the difference in tongue posture attributable to vowel differences, the location and type of contact between the tongue tip and the palate illustrates the problem. With neither linguistic nor other obvious cause, a variety of acts expresses /d/ (see Figure 3).

Identifying the acoustic phonetics of variation in perception also required instrumental methods. Although introspection can be used to recognise allophones through analytical listening, spectrograms were required to discover the acoustic variation inherent in non-allophonic variation. To take a classic instance, the acoustic expression of /d/ is conditional on its vowel context (Liberman et al., 1967). If the vowel following the consonantal release is /i/, there is a rising frequency transition of the second formant; however, if the vowel is /u/, a falling frequency transition is observed after the stop hold is released. There is no noticeable difference in the quality of the consonant; the difference in frequency transition is simply the acoustic pattern occasioned by the perception of /d/ in different vowel environments. In both production and perception of speech, the relation between phoneme and expression is one-to-many, the antithesis of stereotypy. Although the underlying causes of such variation have been elusive to characterise, it is clear from such examples that the variants are not well described by a notion of dispersion about a central tendency.

Freedom from specificity

To appreciate the extent to which the articulated form of a phoneme can vary, it is useful to recall the conditions that produced the descriptions of coarticulatory effects on tongue shape, for instance. The shapes shown in Figure 3 derive from carefully articulated utterances of canonical phonemic form. These are far from typical instances, minimising the competition among expressive functions that use the same set of articulators. In healthy talkers, speech not only manifests the influence of phonemic properties, of course, but also expresses the talker's mood and vitality; the phonetic variants that indicate regional dialect and a talker's idiolect within it; the phonetic properties that signal formality or informality; and, other idiosyncratic states (Remez, 2010). Ordinary articulation is full of instances in which phonemic goals are compromised to allow concomitant aims, and do not mention this to your mother, who told you not to talk with food in your mouth. When this occurs, while sustaining several concurrent expressive aims, articulation must also retain the bolus of food in the oral cavity to prevent inadvertent aspiration and unintended ejection from the front of the mouth. Tongue movement must accordingly comply with this goal while producing a phonemically expressive articulation. It is not known whether speaking with food in the mouth requires creativity, or is an aspect of the potential motor equivalence

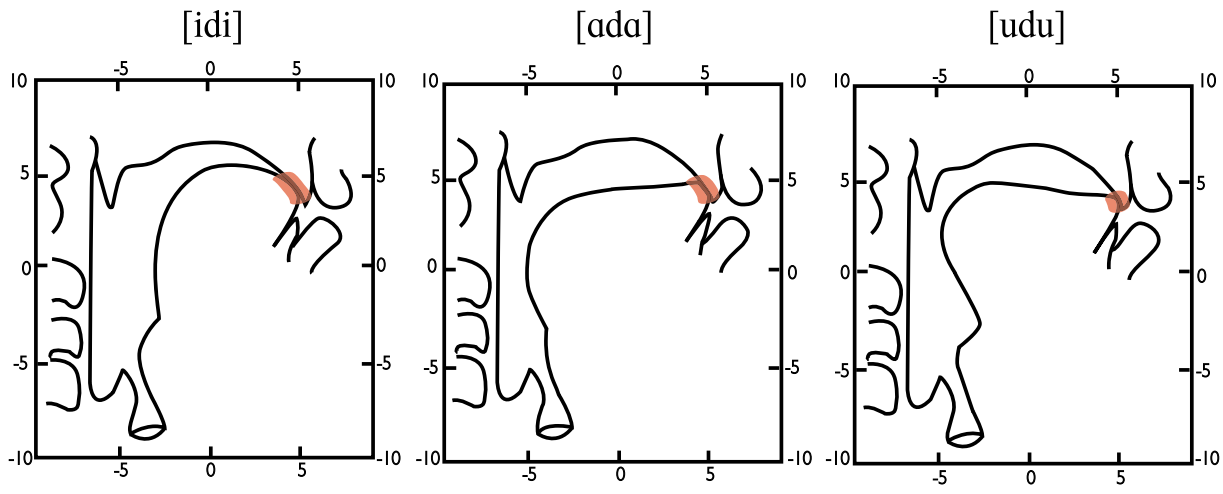

Figure 3. Radiograph tracings of intervocalic /d/ produced in three vowel environment, spoken by and reported by Öhman (1967). The highlights outline the different form of contact articulated in each vowel environment. Reprinted with permission from: Öhman (1967). Copyright 1967, Acoustical Society of America.

inherent in the plans for speech production when a function of suspended deglutition is overlaid.

Remedial interventions also make use of this freedom from specificity when speech is brought to clinical attention, for instance, to manage the hypophonia and dysarthria of Parkinson's disease, or to improve production when articulation has been hampered following clinical attention, as it might be after partial glossectomy. In each of these, the opportunity to produce canonical phonemic form is compromised. Nonetheless, many patients are successful in indicating phoneme contrasts by means of vocalisations that, because of anatomical and physiological changes affecting the tongue, must fail utterly to approximate the tongue shapes shown in Figure 3. The variety of articulatory gambits observed in ordinary and extraordinary conditions of production must itself be convincing that the vocal expression of phonemes does not require the articulation of specific motor acts.

The perceptual analogue of this freedom from specificity was established in the initial era of synthetic speech, during which the means to provide a lifelike replica of speech acoustics exceeded the grasp of technology. Although synthesis approximated the short-term characteristics of speech spectra, intelligibility surpassed naturalness, by far (Liberman & Cooper, 1972). If this outcome had initially seemed like a kind of stimulus generalisation in which an approximation attained the effectiveness of the real thing, experiments continued to challenge this normative assumption about the acoustic elements observed in natural speech – the speech cues (Raphael, 2005). Sine-wave speech disposes of all of the natural products of vocalisation yet perceivers tolerate the contrapuntal whistling carrier, reporting the words and even the nonsense syllables its modulation conveys (Remez, 2008; Remez et al., 1981). Although sine-wave speech does provide a kind of acoustic caricature of formant frequency variation, noiseband-vocoded speech does not. This representation of speech was created to model the likely effects of the electrocochlear implant, imposing sufficient acoustic blur on the speech spectrum that neither individual resonances nor the spectrum of momentary transients are preserved, yet intelligibility is good (Dorman, Loizu, & Rainey, 1997; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995).

Both sine-wave speech and noiseband-vocoded speech use a carrier that is uniform acoustically, and with such simplified spectra the perceptual tolerance of variety might receive only a moderate challenge. The evidence of acoustic chimaeras, then, could be understood to be definitive (Smith, Delgutte, & Oxenham, 2002), with signal elements that can truly be non-uniform in composition and unrelated to speech. This is the consequence of the chimerical combination of two samples, one spoken and the other chosen arbitrarily. The spectrum envelope of the speech signal is estimated and imposed on the zero-crossings of the other, arbitrarily chosen sample. Whether the original samples are exotic or prosaic, the resulting chimerical waveform exhibits a spectrum envelope sufficient to evoke an impression of the original phonetic sequence, and the zero-crossings are adequate to create an impression of the original non-vocal source, be it musical, mechanical or synthetic. The intelligibility of the resulting waveform does not depend on use of zero-crossings with a uniform or stationary spectrum. Both aspects of the chimerical sound are resolvable perceptually, a consequence of the coincidence of short-term spectral properties of the source filtered by the time-varying resonance structure of a speech sample. (Listening examples of these are provided online in Remez, 2008.)

For many years, the attention of researchers has focused on the precise characteristics of discriminable moments in a speech spectrum and their statistical relation to the phonemes. The explanation offered for speech perception in many quarters of the research community continues to sustain an echo of this approach. Nonetheless, the evidence is clear that the perception of speech differs enormously from a function to tally and categorise elements of a closed inventory of canonical acoustic moments. Instead, perception depends on sensitivity to the modulation imposed on a carrier free to vary hugely. In this respect, the freedom from canonical form that derives from the stability of linguistic contrasts establishes an analogy between production and perception.

Points of disanalogy

A talker alone in a room would produce speech for no one else to hear. The parity between production and perception would be perfect, because the variety of spoken forms produced by this hypothetical talker would exactly match the properties resolved perceptually. And, to complete the *Gedanken* experiment, this listener would always know what the audible talker intended to say. The actual social ecology of language differs enormously from this circumstance, and from these conditions disanalogies arise between production and perception. In one obvious way, these reflect a general disanalogy between action and perception. With groceries and Mount Everest in mind, it is clear that the perception of objects and events does not depend on a corresponding capability to produce them. So, in acknowledging a stark asymmetry between production and perception, we can turn to the technical investigations showing that production and perception stand in striking disanalogy in some ways:

(1) perception is talker-contingent;
(2) perception and production are adaptable, but differently;
(3) adaptation in perception does not recalibrate production.

Talker-contingent perception

Perhaps the most obvious disanalogy stems from the heterogeneity of individuals within a community. Adults differ in anatomical scale, and children with immature vocal tracts are the smallest members of any language community. Despite the evident sharing of a lexicon and the cognitive capability for generating phrases, clauses and the rest, each talker's speech is unique, acoustically, a consequence of vocal anatomy and idiosyncrasies of articulation. Indeed, each talker's phonetic expression will vary across alterations of careful and casual speech, formal and informal, standard and vernacular. As a listener, each member of a community recurrently encounters the common stock of linguistic devices, but each instance is specific to the anatomical and phonetic conditions of its creation. Even if the range of variants is limited by restricting consideration of talkers who produce a single coherent dialect, it is still unmistakable that a talker's challenge differs from a listener's. While a talker is responsible for competent production using one vocal apparatus only, a listener's perceptual challenge is to apprehend the linguistic attributes expressed by many varied vocalising individuals. In this condition, it is not surprising to find that cognitive resources are tuned during perception to the contingencies of recognition of the speech of an individual talker.

A landmark study that examined talker-contingent perception used two procedures to estimate the perceptual effects (Nygaard, Sommers, & Pisoni, 1994). First, a training paradigm was used to establish familiarity with a set of talkers among a group of listeners. On each trial of a test session, a monosyllable word was presented to a listener who was asked to identify which of the ten talkers had produced it. Second, to create an assay of the perceptual consequences of familiarity, a test of word recognition was conducted. Although the words composing the second test had not been used in the talker learning trials, some had been produced by the talkers whom the listeners had just learned to identify. Performance in recognising words was better for samples spoken by familiar talkers than those produced by unfamiliar talkers; this performance level difference was observed at all signal-to-noise levels (see Figure 4). Because the word sets used in training talker identification did not include those used for testing spoken word recognition, an instance-specific benefit of the kind ascribed to implicit learning could not be responsible for the relative immunity

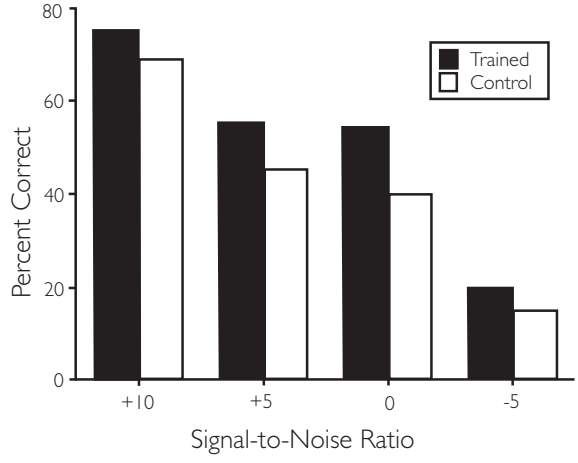

Figure 4. Performance on a task of spoken word recognition presented at four signal-to-noise ratios (measures from Nygaard et al., 1994). In all conditions of presentation, recognition of words spoken by familiar talkers exceeded recognition of words spoken by unfamiliar talkers. See text for a complete description.

to noise brought about by familiarity with the talker. What did produce the benefit?

It seems as though the familiarity that facilitates the resolution of linguistic properties has two components: one is auditory and the other phonetic. This conclusion is suggested by a two-part study of sine-wave spoken word identification (Remez et al., 2011). In an exposure phase, a group of sentences was presented to listeners to transcribe; in a second part, a set of monosyllable sine-wave words was presented for identification. The exposure phase varied the attribute that matched the sine-wave words. Idiolect and acoustic spectrum matched when the sine-wave sentences were based on speech samples of the talker who had also provided natural models for the sine-wave words. When sine-wave sentences based on a different talker were used, the acoustic spectrum matched but idiolect mismatched. When natural sentences were used that had been produced by the talker who also provided the natural models for the sine-wave words, idiolect matched but the spectrum mismatched. In a control, some listeners took the test of word recognition without any prior exposure to the talker. The outcome was clear. Facilitation of sine-wave word identification occurred only when the spectrum and the idiolect were the same in the exposure sentences and in the words. While this evidence of conjunction must yet be submitted to stronger test, it does appear that perception is susceptible to aspects of the auditory quality of a talker's spectrum as well as the distinctive phonetic aspects of the talker's expression: the idiolect. This functional tuning of perception to individual talkers has no counterpart in production and is a point of disanalogy between expressive and receptive language.

Adaptive differences in production and perception

Production and perception of speech are adaptive functions, though not in the technical sense that was common in psychology until recently. *Adaptation* in this older technical sense originates in the adjustments of visual sensitivity across a wide range of illumination (Boring, 1942) and is observed chiefly as changes in threshold. A newer sense deriving from consideration of adaptive systems of great complexity has supplanted this older meaning, and now pertains to the systematic adjustment to local conditions that preserves goal-directed stability. It is in that newer sense that production and perception of speech are adaptive. However, their adaptive characteristics differ greatly, both in the adaptable properties and in the time-course.

Production varies in precision, pacing the changes in communicative risk. A prominent example of this was reported by Lieberman (1963), who found that the articulatory care applied to the word NINE differed with its cloze predictability. When it was predictable from context ("A stitch in time saves NINE".), it was less robustly realised than when it was produced in a context in which it was unpredictable ("The word that you will hear is NINE".). A talker arguably does such things because of a disposition to minimise productive costs at the expense of the listener, if this can occur without getting caught (Lindblom, 1990). When perception becomes too effortful due to a talker's excessive thrift, a listener might insist on repetition and clarification, escalating the talker's cost catastrophically in this view of conversation as an arms race.

In fact, the phenomenon described by Lieberman (1963), if not exactly false, is actually less true than the original evidence and common sense dictate. This is shown in a recent study (Clopper & Pierrehumbert, 2008). In this version of the paradigm, the context of occurrence of target words was manipulated to vary in predictability, once again, but the target words were also chosen with a crossing indexical property. Due to the Northern Cities Chain Shift (Labov, 1998), the target words offered juicy opportunities for Midwestern talkers to express a salient dialectal marker of identity. In the sentence pair, "Please wipe your feet on the MAT", and "Peter considered the MAT", for instance, the target word might exhibit /æ/-raising to mark the dialect, or articulatory expression might drift toward the canonical form. The outcome depends on a tussle between phonemic and indexical motives, each of which promotes a different and mutually incompatible phonetic form. Given the choice between communicating effectively and marking identity phonetically, talkers mainly reserved the regional expression for the more predictable cases and offered a version of MAT close to the canonical phonemic form when it was less predictable from context. Or, to be more precise, they did this if they lived north of Interstate-70. Talkers from the Midlands, south of I-70, were indifferent to the opportunity to adapt the production of vowel height in order to mitigate communicative risk. Articulation of indexical phonetic features consistently suppressed canonical phonemic expression in those talkers.

In these two cases, adaptive changes were driven presumably by a talker's a priori estimate of the perceptual ease or difficulty for a listener to cope with the habitual conditions of production. The changes that were reported are attributable to an aspect of the talker's knowledge of language and implicit empathy with the listener. But, other conditions also drive adaptation by causing a change in the mechanics of production. In a report about bite-block speech (Fowler & Turvey, 1980; cf. Baum, Kim, & Katz, 1997), excursions of the mandible were restricted during the production of vowels by the assignment to retain a 10 mm by 14 mm wood dowel between the teeth. Although American English vowels are said to differ in height, this method fixes the height of the jaw, forcing an adaptation to the constraining condition. While the specific

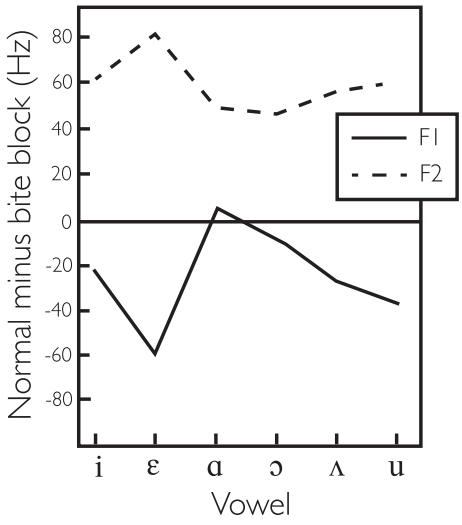

Figure 5. Acoustic measures of six vowels showing the differences between unconstrained (jaw free) and bite-block (jaw fixed) production. Both formant measures revealed that compensation for this perturbation of articulation occurred swiftly and with negligible acoustic consequences (measures reported by Fowler and Turvey, 1980).

acts of articulation differed in jaw-free and jaw-fixed conditions, the subjects hardly differed in the acoustic effects of the bite-block relative to the unconstrained production and were highly accurate in compensating in the lowest resonance typically associated with the opening and closing of the jaw (see Figure 5). Although the condition that drove adaptation was physical rather than conceptual, compensation was instantaneous, or nearly so.

The jaw is a mobile articulator, ordinarily, and controlling its height must be a significant component of speech production. Because the position of the jaw relative to the rest of the vocal tract is changing constantly during an utterance, perhaps it is not altogether unexpected that constraints on its movement are readily adaptable. The jaw is being moved and tracked, anyway, in this view. A relation which is less compliant to adaptation can be found in the link between sound production and auditory reafference. Although the lag of auditory reafference in speech production is too great to allow monitored regulation of the sequential production of phonetic segments, which can approach 20 Hz, other aspects of reafferent control are available at longer lag. Because this link between perceptual sensitivity and productive control is generally stable, especially with regard to acoustic frequency, it is adaptable, but only slowly. It took 20 minutes of exposure (Houde & Jordan, 1998) for subjects to recalibrate to an artificial perturbation in formant frequency; and, it took 70 trials for subjects to recalibrate to perturbed reafference about phonatory frequency (Jones & Munhall, 2000). In each of these cases, the vocal sound produced by a subject was submitted for fast alteration by digital signal processing and returned through headphones. The perturbation was gradual and drifting, and accordingly not noticed by the subject, who took the altered reafference as if it were veridical. At some point, the perturbation was sufficient in magnitude to elicit compensation in production, in which a subject raised the jaw to oppose the sensory effect of a perturbation that raised the frequency of the first formant; or, a subject raised the frequency of phonation to compensate for reafference that was lowered in frequency. The inherent stability of this link in the talker's experience plausibly confers some immunity to adaptation, observed here as a lag in the time-course of compensation.

The difference between stability and lability in adaptation of speech production also depends on the anatomical structure involved. For instance, the hard palate is a rigid and fixed structure that is very nearly consistent in shape from childhood to adulthood (Boë, et al., 2006). The introduction of a pseudopalate can be used to disrupt the topography of this oral surface (Baum & McFarland, 1997), perturbing the articulatory geometry of phonemes with alveolar and post-alveolar phonetic place (see Figure 6). In this circumstance, the compensation for the change in shape of this typically non-deformable surface is not immediate, and talkers took 20 minutes of exposure to adapt. However, in order to produce a fully natural sounding production, several days to several weeks of exposure might have been required (cf. Hamlet, Stone, & McCarty, 1978). In a perturbation study with another hard tissue, Jones and Munhall (2003) outfitted a subject with a dental appliance that lengthened the upper incisors, altering the aerodynamic requirements for producing /s/. Adaptation was not immediate, and subjects took about a half hour of exposure to adjust. Principally, the recalibration involved a change in tongue shape and placement to direct a jet of air at the back of the lengthened teeth, thereby producing the consonant. To understand why the course of adaptation is immediate in one circumstance and lagging in others, recall that some structures vary in position and motion constantly, while others are fixed in shape and position within which the play of mobile articulators occurs. The stability of production inheres, presumably, from the devotion of adaptable resources to the changeable properties, relegating the fixed properties of the vocal tract to a status that is relatively unmonitored, at least with respect to adaptive tracking and dynamic recalibration. The perception of speech contends with a rather different condition.

The physical uniqueness of each talker imposes an organic effect on vocal sound, as if the functional expression of linguistic properties rides on an anatomically based acoustic signature of the talker. Such scale variation might be normalised, perceptually (Pisoni, 1997), but other functional aspects unique to each talker are irreducible: age, sex, gender, mood, vitality, dialect, accent, idiolect and idiosyncrasies of anatomy all contribute to the phonetic repertoire deployed expressively. There

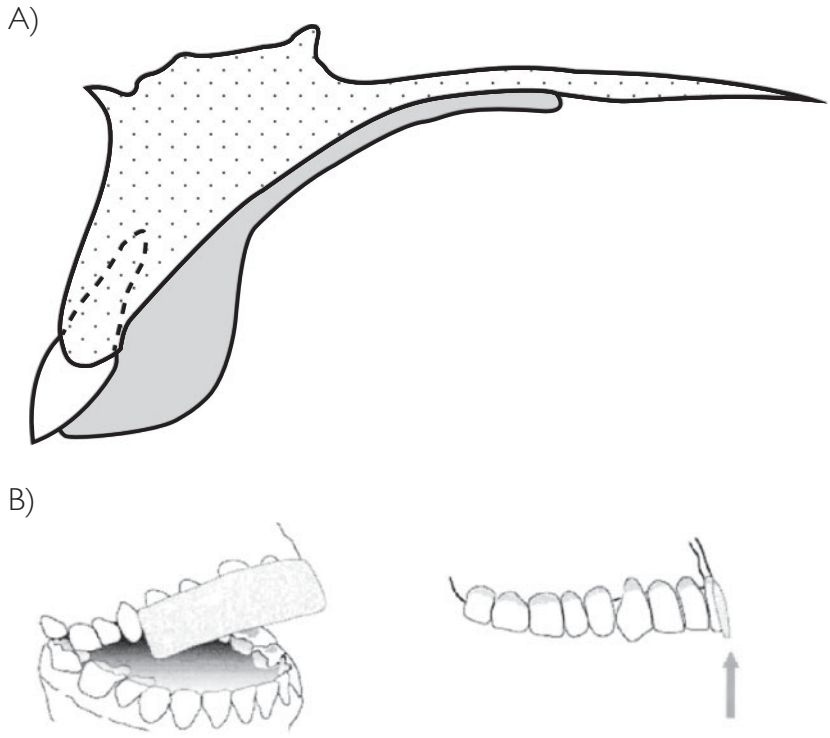

Figure 6. (A) A sagittal view of the placement and shape of a pseudopalate used to perturb the topography of articulation. Reprinted with permission from: Baum and McFarland (1997). Copyright 1997, Acoustical Society of America. (B) A dental appliance used to perturb the production of /s/. The left panel shows the upper and lower teeth; the right panel shows a sagittal view depicting the effective lengthening of the lingual surface of the upper central incisors. Reprinted with permission from: Jones and Munhall (2003). Copyright 2003, Acoustical Society of America.

are few estimates of the size of the set of talkers with which a single individual is familiar, but it is safe to speculate that in our era of mediated social interchange it is large. Some tests demonstrate impressive familiarity with talkers who are merely famous, and who do not belong to the intimate social network of the subjects (Van Lancker, Kreiman, & Emmorey, 1985). In this kind of environment, a talker and a listener who share a language are aligned on other properties only accidentally, and these offer the opportunities for adaptive recalibration in perception. The inevitability of such functions fits the fluency with which they occur.

How rapidly does a perceiver adapt to a new talker? If synthetic speech can be mentioned, a single sentence of six syllables is all that is required to fine-tune an analytic standard for perception (Ladefoged & Broadbent, 1957). Because vocal impressions of a talker also accompany the perception of sine-wave speech, it is sensible that this finding of rapid establishment of talker-contingency applies equally to that kind of impossible spectrum (Remez, Rubin, Nygaard, & Howell, 1987). In each of these cases, the listeners spoke the same language and roughly the same dialect as the natural models for the synthesis. Listening to foreign-accented speech is more demanding, perhaps, though hardly less fluent in adaptation. A study of foreign-accented speech with closely related (English-Spanish) and remotely related (English-Chinese) first languages showed very fast tuning to a talker's individual characteristics (Clarke & Garrett, 2004), within the span of two sentences.

Each of these perceptual studies used procedures to focus a participant's attention on linguistic properties and estimated the exposure to a talker's unique characteristics that enhanced word recognition. The speech samples were unrestricted, phonemically, and by the classic standard of Pollack, Pickett, & Sumby (1954) – ≥8 segments – provided a phonemically varied utterance sufficient to evoke an impression of a specific talker. Yet, these studies overestimate the phonemic variety required to establish a new perceptual impression of a talker. Astonishingly, in a study of a single feature contrast, listeners formed impressions of different individuals whose identity depended solely on the voicing of American English stop consonants present in four words: DOWN, TOWN, DIME and TIME (Allen & Miller, 2004). Beginning with natural samples of two talkers, a pitch-synchronous linear-prediction vocoder was used to synthesise voicing variants for each, nominally, Annie and Laura. Training procedures taught participants to label individual samples as spoken by one or the other talker, and listeners were able to

generalise with novel items that were intermediate between the training items. The generalisation performance indicated that listeners had induced the identities of two talkers, one of whom had characteristically long voiceless voicing onset times (VOTs) and the other shorter VOTs. The acoustic contrast between syllables was subtle but evidently learnable. The training items for one talker exhibited a +voiced VOT of 13 ms and a –voiced VOT of 172 or 182 ms; the other was given a +voiced VOT of 15 ms and a –voiced VOT of 78 or 87 ms. More remarkable is that the talker training trials were interleaved, demanding that perceivers track the idiolects of two talkers concurrently. Although these findings are based on a formal training and generalisation procedure, an analysis of the warm-up interval when subjects were simply becoming familiar with the button pad used to collect the responses revealed that within about six syllables of exposure to each voicing contrast the participants were already forming impressions of Annie and Laura (Miller, 2005). In contrast to the long exposure required for adaptation in the production of speech, the perception of speech is fast and supple, and this is a prominent disanalogy between the two aspects of expression.

Perception does not recalibrate production

The first disanalogy between production and perception discussed here is a result of intrinsic differences in the variety of conditions that each function must contend with. Production using a single vocal apparatus is inherently a narrower task than perception of speech originating in many talkers, or, to get the count right, any talker. The second disanalogy pertained to the interchange between compliance and resistance to change. Although both production and perception are adaptable to the conditions of expression, far greater compliance of perception is observed both in range and in time-course of adaptation. But, this consideration examined the expressive and receptive functions in parallel. The production and perception of speech alike are driven by the knowledge of phoneme contrasts and their phonetic expression, and it is reasonable to suppose that adaptation in one function would affect the other. But, we should not be satisfied with obvious or truistic cases, for instance, that learning to identify a new word promotes its production. Instead, the disanalogy is well evident considering the instances in which perception recalibrates production at the level of fine phonetic detail.

Some empirical proofs exist and have been interpreted as evidence of perceptual recalibration crossing over to production (for instance, Cooper & Lauritsen, 1974). The best known of these is a series of experiments on vocal repetition of words (Goldinger, 1998). A subject was asked to listen to a recorded utterance of a word and to repeat it, either immediately, or after a brief delay was imposed. The repetition was recorded and used in a subsequent perceptual test comparing the elicited speech samples to the eliciting samples. A new set of listeners judged the similarity of the eliciting and the elicited utterances, and their reports in aggregate showed that words spoken immediately after an eliciting utterance were more similar to the eliciting speech than were utterances produced after a brief delay. Every perceptual trace fades, but while it is still fresh, it seems as if the perceived form of a word influences the subsequently produced form, evidence of analogy. However, even if elicited utterances were more similar under one condition than another, they each still exhibited the largely unaltered characteristics of the talker who produced it. In other words, despite a marginal increase in similarity under some eliciting conditions, none of the elicited utterances was ever a remotely faithful replica of an eliciting utterance. And, the similarity was greater for rare words than for common words, warranting a specific constraint on interpretation of the evidence: Perception, a talker-contingent function, is unlikely to force a radical and global talker-contingent recalibration of production. Ultimately, the better portion of production is not available for assimilation to perceived speech, a consequence of disanalogy in their dynamic.

A similar push-and-pull between production and perception is also observed across language (Sancier & Fowler, 1997). This acoustic phonetic study examined the production of voicing by a bilingual individual after short-term immersion in the native and second language environments. The first language, Brazilian Portuguese, differs from the second language, American English, in timing and aspiration of voicing. For the bilingual talker in this project, expression of two voiceless stop consonants in each language drifted slightly in the direction of the hypothetical norms of the local language, as measured from canonical expressions. This assimilation might have been offered as evidence that perception recalibrated production. However, the extent of recalibration was small and imperceptible if consistent in the second language. Overall, production remained fundamentally stable, despite these effects of exposure.

Evidence of assimilation amid productive stability is also available from experiments using conversational settings, assuring the naturalness of the investigative methods. In a project using a route-tracing task to elicit speech, a conversing dyad was recorded over the course of a bout (Pardo, 2006). Because the landmarks on each partner's map differed from the other's, there was a lot to discuss. One member of each pair gave instructions to the other, who drew the route. The recorded speech of the dyad during the map task was inspected to identify recurring lexical items, and these items included instances that were spoken before, during, and after the task. Exposure to each other's speech promoted the perception

of talker-specific characteristics, and over the course of the task the some conversational partners also became more similar in their production. This was established by extensive testing that compared the similarity of utterances of the dyads from different phases of a session. However, the resistance to assimilate, productively, was prominent in the measures. For example, female dyads were unlikely to assimilate to each other, and assimilation in the male dyads was not reciprocal. That is, the senders of the routes converged in speech to the receivers, but not vice versa. Yet, as Goldinger had observed, the increases in similarity were subtle; there was little evidence of mimicry or outright imitation. Instead, small adjustments were arguably responsible for small changes in similarity, and even an utterance exhibiting convergence, phonetically, remained characteristic of the talker who spoke it.

The propensity to converge at all is evidence that perception of another's speech can leak into one's own production. Nonetheless, individual habits of articulation are apparently stable over decades of adult exposure to new dialects and idiolects (House & Stevens, 1999). Adopting a broad perspective, we must acknowledge that there are significant brakes on convergence that prevent us from talking alike no matter how much time we spend together. These phenomena are not well understood, technically, and the burgeoning field of phonetic assimilation research promises to expose this complex and intriguing link between production and perception. In broad strokes, though, the functions are hardly baffling.

The production of speech serves linguistic and indexical aims. By formulating semantic intentions and expressing them, a talker denotes a message, but because the form of expression, phonetically, bears the anatomical, affective, dialectal and idiolectal attributes of the talker, the speech that issues from the mouth is multiplexed. In addition to articulating the phoneme contrasts adequately, each talker offers an assortment of features that indicates membership in a group and includes a few features to remain uniquely identifiable within the group. These opposing tendencies to assimilate and to resist assimilation socially (Pickett & Brewer, 2001) provide the conditions that apparently energise this disanalogy between production and perception. Speaking in one's own voice obliges resistance to adopt the entire variety of phonetic variants encountered in the playground or on the boulevard, despite the evident importance of resolving those variants in perceiving utterances. Research on the properties of foreign accent that preserve the phonological commitments of the first language also shows that there are cascading developmental challenges to this disanalogy of production and perception. Even when motivation is great, aspects of production can remain unassailable despite perceptual acuity (though, see Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997).

Conclusion

In this review of classic and recent perspectives on the system architecture of speech, an intimate connection between production and perception is presumed by axiom and defended by evidence. Buttressed in this manner, the reciprocal functions that have absorbed so much useful attention in our community are resistant to facile disproof. It would be astonishing, indeed, for a study to appear in a technical journal next year showing that production and perception are actually unrelated linguistically, cognitively, or neurophysiologically. The relatedness of production and perception are a portion of what we mean by acknowledging that language is a medium of representation.

Yet, the overall description of the relation presented here is one of fundamental disanalogy, a conclusion that acknowledges vastly different operating characteristics and control parameters in production and perception. Although production and perception distinguish words by the same abstract phonemic contrasts, perception must accommodate far wider phonetic variation in expressed forms than production. This is simply due to the vastly greater variation inherent in speech perceived by a single individual, in contrast to the speech produced by a single individual. Perception contends with the speech of any talker; the dynamics of the control of production must simply produce consistency in the speech of a single talker. The variety of phonemic features and expressed forms simply differs in the two modalities, casting them in fundamental disanalogy despite shared properties.

Because the points of disanalogy seem irreducible, the description of language function that results is itself heterogeneous rather than uniform. A collection of diverse resources is apparently operating to establish and maintain the expressive functions of spoken language. Their coordinated action permits production and perception to converge linguistically, personally and circumstantially despite constituent functions that differ intrinsically. In order to develop a more refined description of the dynamic that implements these complementary aspects of expression, several questions are pressing:

- How do linguistically competent adults tolerate such enormous phonetic variation?
- How does production in childhood reconcile the opposing pressures of alignment with the community and idiolect creation?
- Can we identify principles distinguishing an expressive dynamic in which *anything goes* from one in which *many things go*?

Acknowledgements

The author is grateful to the organisers of IWOLP-7 for the opportunity to talk and to write about this topic; to Michele Miozzo for insisting that the viewpoint of a perceptionist would be welcome at a production meeting; and, to David Pisoni for his

counsel about intellectual heritage. For preparing the figures, I offer 1×10^6 thanks to Rebecca Giglio. And, Figure 1 could not have been made without the astute pen of Phoebe Fisher, the sound advice of Emily Thomas and the good work of Andrea Wycoff, Kate Francis, Sarah Alice Hanna, Colin Simon-Fellowes and Liam Simon-Fellowes.

Funding

This research was supported by the National Institute on Deafness and Other Communication Disorders under grant no. [DC000308].

References

Abbs, J. (1973). The influence of the gamma motor system on jaw movements during speech: A theoretical framework and some preliminary observations. *Journal of Speech and Hearing Research*, *16*, 175–200. doi:10.1044/jshr.1602.175

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice onset-time. *Journal of the Acoustical Society of America*, *115*, 3171–3183. doi:10.1121/1.3106131

Baum, S. R., Kim, J. A., & Katz, W. F. (1997). Compensation for jaw fixation by aphasic patients. *Brain and Language*, *56*, 354–376. doi:10.1006/brln.1997.1734

Baum, S. R., & McFarland, D. H. (1997). The development of speech adaptation to an artificial palate. *Journal of the Acoustical Society of America*, *102*, 2353–2359. doi:10.1121/1.429429

Boë, J.-L., Granat, J., Badin, P., Autesserre, D., Pochic, D., Zga, N., … Ménard, L. (2006). Skull and vocal tract growth from newborn to adult. *Proceedings of the 7th International Seminar on Speech Production*, 75–82. Ubatuba, Brazil: ISSP.

Borden, G. J., Harris, K. S., & Oliver, W. (1973). Oral feedback, I. Variability of the effect of nerve block anesthesia upon speech. *Journal of Phonetics*, *1*, 289–295.

Boring, E. G. (1942). *Sensation and perception in the history of experimental psychology*. New York, NY: Appleton-Century.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*, 2304–2310. doi:10.1121/1.418276

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 115–127. doi:10.1037/0096-1523.27.1.115

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, 116, 3647–3658. doi:10.1121/1.1815131

Clopper, C. G., & Pierrehumbert, J. B. (2008). Effects of semantic predictability and regional dialect on vowel space reduction. *Journal of the Acoustical Society of America*, *124*, 1682–1688. doi:10.1121/1.2953322

Cooper, W. E., & Lauritsen, M. R. (1974). Feature processing in the perception and production of speech. *Nature*, *252*, 121–123. doi:10.1038/252121a0

Denes, P. B., & Pinson, E. N. (1963). *The speech chain*. New York, NY: Bell Telephone Laboratories [distributor: Williams & Wilkins Co., Baltimore].

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America*, *102*, 2403–2411. doi:10.1121/1.419603

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fowler, C. A., & Turvey, M. T. (1980). Immediate compensation in bite-block speech. *Phonetica*, *37*, 306–326. doi:10.1159/000260000

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279. doi:10.1037/0033-295X.105.2.251

Hamlet, S., Stone, M., & McCarty, T. (1978). Conditioning dentures viewed from the standpoint of speech adaptation. *Journal of Prosthetic Dentistry*, *40*, 227–248. doi:10.1016/0022-3913(78)90160-9

Houde, J. F., & Jordan, M. (1998). Sensorimotor adaptation in speech production. *Science*, *279*, 1213–1216. doi:10.1126/science.279.5354.1213

House, A. S., & Stevens, K. N. (1999). *A longitudinal study of speech production, I: General findings*. Speech Communication Group Working Papers, XI, 21–41. Cambridge, MA: RLE, MIT.

Jakobson, R., & Halle, M. (1956). *Fundamentals of language*. The Hague: Mouton.

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of *F*0 production: Evidence from feedback perturbation. *Journal of the Acoustical Society of America*, *108*, 1246–1251. doi:10.1121/1.1288414

Jones, J. A., & Munhall, K. G. (2003). Learning to produce speech with an altered vocal tract: The role of auditory feedback. *Journal of the Acoustical Society of America*, *113*, 532–543. doi:10.1121/1.1529670

Labov, W. (1998). The three dialects of English. In M. D. Linn (Ed.), *Handbook of dialects and language variation* (pp. 39–81). San Diego, CA: Academic Press.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*, 98–104. doi:10.1121/1.1908694

Lane, H. (1968). On the necessity of distinguishing between speaking and listening. *Studies in language and language behavior, progress report VI*, 22–49. Ann Arbor, MI: University of Michigan.

Lane, H. L., Catania, A. C., & Stevens, S. S. (1961). Voice level: Autophonic scale, perceived loudness and effects of sidetone. *Journal of the Acoustical Society of America*, *33*, 160–167. doi:10.1121/1.1908608

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–131). New York, NY: Wiley.

Liberman, A. M., & Cooper, F. S. (1972). In search of the acoustic cues. In A. Valdman (Ed.), *Papers in linguistics and phonetics to the memory of Pierre Delattre* (pp. 329–338). The Hague: Mouton.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 421–461. doi:10.1037/h0020279

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36. doi:10.1016/0010-0277(85)90021-6

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*, 172–187.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H & H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht: Kluwer.

Massaro, D. W. (1994). Psychological aspects of speech perception: Implications for research and theory. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 219–263). San Diego, CA: Academic Press.

Mattingly, I. G., & Liberman, A. M. (1988) Specialized perceiving systems for speech and other biologically significant sounds. In G. M. Edelman, W. E. Gall, & W. M. Cowan (Eds.), *Auditory function: Neurobiological bases of hearing* (pp. 775–793). New York, NY: Wiley.

Miller, J. L. (2005). Listener sensitivity to fine phonetic detail in speech perception. *Minutes of the Columbia University Seminar on Language and Cognition, 2005–2006* (pp. 56–76). New York, NY: Columbia University.

Moll, K. L. (1960). Cinefluorographic techniques in speech research. *Journal of Speech and Hearing Research*, *3*, 227–241. doi:10.1044/jshr.0303.227

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*, 42–46. doi:10.1111/j.1467-9280.1994.tb00612.x

Öhman, S. E. G. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America*, *41*, 310–320. doi:10.1121/1.1910340

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*, 2382–2393. doi:10.1121/1.2178720

Pickett, C. L., & Brewer, M. B. (2001). Assimilation and differentiation needs as motivational determinants of perceived ingroup and outgroup homogeneity. *Journal of Experimental Social Psychology*, *37*, 341–348. doi:10.1006/jesp.2000.1469

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–32). San Diego, CA: Academic Press.

Pollack, I., Pickett, J. M., & Sumby, W. H. (1954). On the identification of speakers by voice. *Journal of the Acoustical Society of America*, *26*, 403–406. doi:10.1121/1.1907349

Quine, W. V. O. (1968). Ontological relativity. *Journal of Philosophy*, *65*, 185–212. doi:10.2307/2024305

Raphael, L. J. (2005). Acoustic cues to the perception of segmental phonemes. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 182–206). Oxford: Blackwell.

Remez, R. E. (2008). Sine-wave speech. *Scholarpedia*, *3*, 2394. doi:10.4249/scholarpedia.2394

Remez, R. E. (2010). Spoken expression of individual identity and the listener. In E. Morsella (Ed.), *Expressing oneself/expressing one's self: Communication, cognition, language, and identity* (pp. 167–181). New York, NY: Psychology Press.

Remez, R. E., Dubowski, K. R., Broder, R. S., Davids, M. L., Grossman, Y. S., Moskalenko, M., … Hasbun, S. M. (2011). Auditory-phonetic projection and lexical structure in the recognition of sine-wave words. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 968–977. doi:10.1037/a0020734

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 41–60. doi:10.1037/0096-1523.13.1.40

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*, 947–950. doi:10.1126/science.7233191

Remez, R. E., & Thomas, E. F. (2013). Early recognition of speech. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*, 213–223. doi:10.1002/wcs.1213

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*, 421–436. doi:10.1006/jpho.1997.0051

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304. doi:10.1126/science.270.5234.303

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*, 87–90. doi:10.1038/416087a

Subtelny, J. D., Pruzansky, S., & Subtelny, J. (1957). The application of roentgenography in the study of speech. In L. Kaiser (Ed.), *Manual of phonetics* (pp. 166–179). Amsterdam: North Holland.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters, Part 1: Recognition of backward voices. *Journal of Phonetics*, *13*, 19–38.

Zevin, J. D., Yang, J. F., Skipper, J. I., & McCandliss, B. D. (2010). Domain general change detection accounts for "dishabituation" effects in temporal-parietal regions in fMRI studies of speech perception. *Journal of Neuroscience*, *30*, 1110–1117. doi:10.1523/JNEUROSCI.4599-09.2010