Published in final edited form as:

Lang Cogn Neurosci. 2018 ; 33(9): 1083–1091. doi:10.1080/23273798.2018.1442580.

SHORT-TERM PERCEPTUAL TUNING TO TALKER CHARACTERISTICS

Robert E. Remez¹, Emily F. Thomas¹, Aislinn T. Crank¹, Katrina B. Kostro¹, Chloe B. Cheimets¹, and Jennifer S. Pardo²

¹Program in Neuroscience & Behavior, Department of Psychology Barnard College, Columbia University

²Department of Psychology Montclair State University

Abstract

When a listener encounters an unfamiliar talker, the ensuing perceptual accommodation to the unique characteristics of the talker has two aspects: (1) the listener assesses acoustic characteristics of speech to resolve the properties of the talker's sound production; and, (2) the listener appraises the talker's idiolect, subphonemic phonetic properties that compose the finest grain of linguistic production. A new study controlled a listener's exposure to determine whether the perceptual benefit rests on specific segmental experience. Effects of sentence exposure were measured using a spoken word identification task of Easy words (likely words drawn from sparse neighborhoods of less likely words) and Hard words (less likely words drawn from dense neighborhoods of more likely words). Recognition of words was facilitated by exposure to *voiced obstruent* consonants. Overall, these findings indicate that talker-specific perceptual tuning might depend more on exposure to phonemically marked consonants than to exposure distributed across the phoneme inventory.

Keywords

speech perception; talker familiarity; perceptual tuning

1. Introduction

Speech perception is a talker-contingent cognitive function. In recognizing spoken words, there is an interval, arguably brief, during which a listener's perceptual resources adjust to the idiosyncrasies of a specific talker. The consequences cut both ways, boosting intelligibility when the talker is familiar, and hampering recognition when the talker is unfamiliar. In the laboratory, even the identification of an isolated nonsense syllable is undermined when the talker can be varied unpredictably from trial to trial in a test procedure (Mullenix & Pisoni, 1990). This conceptualization of speech perception emerged in the past two decades (e. g., Allen & Miller, 2004; Goldinger, Pisoni & Logan, 1991; Kraljic, Brennan & Samuel, 2008; Nygaard, Sommers & Pisoni, 1994), and places a specific

qualification on the familiar description of invariance and variability which has colored so much discussion of the challenge of recognition.

1.1. Idiolect and indexical attributes

Whether words, or syllables, or individual phonemes, linguistic forms are shared by the talkers of a language. However, this common stock of linguistic items is only ever expressed in personal form; and, each utterance is circumstantially unique, an expression of variability that obscures the linguistic properties in speech spectra. Classical perceptual studies had chiefly concerned the effects of coarticulation as a cause of the departure from invariance in the correspondence of linguistic form and acoustic form. In consequence, perceptual functions were said to apply an inverse function principally to undo coarticulation, thereby unmasking the discrete phonemes in the acoustic signal (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967). In this conceptualization, perception of speech is a function devoted to the identification of phonemes by indifference to the subphonemic variation due to coarticulation. However, more recent studies have viewed production of speech as a multiplex bottleneck in which the acoustic phonetic form reflects a talker's age, vitality, affect, motivation, regional origin, idiolect, stylistic habits and history of dental work in addition to canonical phonemes (reviewed by Remez & Thomas, 2013). From this perspective, coarticulation is a single cause among many that drive phonetic form to depart from invariant phonemic expression. A perceptual function that applied to acoustic or articulatory phonetic form solely to undo the obscuring effect of coarticulation would leave unaffected the merger of phoneme expression with all of the personal, circumstantial features expressed concurrently through vocalization. Inasmuch as perceivers are acutely aware of these dimensions of variation among interlocutors, it seems implausible that the perception of phoneme contrasts includes an early function that strips the subphonemic variation from sensory samples of speech (cf. Johnson & Mullenix, 1997).

Indeed, evidence has shown that a talker is readily identifiable from phonetic expression (Remez, Fellowes & Rubin, 1997). The subphonemic within-category phonetic variation which once was believed to be inaudible — due to categorical perception — is instead a rich source of attributes affecting the perceptual identification of individuals. Such indexical properties were believed at one time to constitute a second message, distinct from the linguistic form, because these were understood as graded aspects of variation in anatomy or posture which caused global spectral effects (Bricker & Pruzansky, 1976; Nolan, 1983; see, also Abercrombie, 1967; Ladefoged, 1967). Instead, the phonetic idiosyncrasies of individuals also apparently function as indexical properties despite their linguistic grain and governance, because these are consistent enough in expression, noticeable, and memorable, and thereby can be used to distinguish one talker from another (Remez, 2010).

The subphonemic phonetic details of speech that function indexically also are potentially responsible for the facilitation of intelligibility when a listener has an opportunity to become familiar with a specific individual talker (Nygaard et al., 1994). For instance, after listeners had learned to identify a small set of unfamiliar talkers, the novel words spoken by these newly familiar individuals were more recognizable than control words spoken by strangers. This advantage was observed at four S/N ranging from +10 dB to -5 dB. Through short-

term perceptual tuning to the subphonemic variants used by a specific talker, due to dialect or idiolect or both, a listener becomes more sensitive to the incidence of idiosyncratic variants when encountering new utterances by the familiar talker.

1.2. A precedent project with Easy and Hard Word Recognition

A recent study of ours (Remez, Dubowski, Broder, Davids, Grossman, Moskalenko, Pardo & Hasbun, 2011) took advantage of the enhancement in intelligibility attributable to exposure to an unfamiliar talker, in an attempt to characterize the crucial dimensions of short-term perceptual tuning to talker characteristics. In comparison to studies that had focused on exposure to a specific phonemic contrast (Allen & Miller, 2004) and to projects that had trained listeners to become familiar with a set of individual talkers (Nygaard et al., 1994), our project simply matched or mismatched the talker attributes common to an exposure interval and a test condition. The rationale was straightforward. If a property of exposure is essential to boost intelligibility, then that property must be shared by exposure utterances and test utterances to induce and then to reveal the benefit. Alternatively, if an aspect of a talker's speech is irrelevant to the perceptual tuning that enhances recognition, then mismatching that attribute between the exposure interval and the test conditions should occur without harming performance. The dimensions in that study (Remez et al., 2011a) which were matched or mismatched included acoustic form and idiolect.

To create the exposure conditions, Remez et al. (2011a) used speech samples of sentences typical in intelligibility testing; these were collected from two talkers. One of the talkers also produced 148 monosyllable words composing the Easy/Hard word lists (Bradlow & Pisoni, 1999). Sentences were used in an exposure condition, while the Easy/Hard word lists were used after exposure in an open set test of spoken word recognition.

Due to inhomogeneities in the distribution of words within the English lexicon, the Easy/Hard word lists can offer a sensitive measure of phonetic acuity and, in this case, the enhancement of phonetic sensitivity following exposure to a new talker. The key notion is the lexical neighborhood, or the number and characteristics of the words that are similar to a given word. According to an estimate of its frequency of occurrence, an Easy word was more common than the similar words composing its lexical neighborhood; it had relatively few lexical neighbors; and, these neighbors were all less common than the Easy word itself. In contrast, a Hard word was less frequent than the words composing its lexical neighborhood; it had relatively more neighbors than an Easy word; and, all of its neighbors were more common than the word itself. Although many aspects of spoken word recognition are signal contingent, and vary as a function of the resolution of acoustic structure, intelligibility is also affected by this kind of signal-independent inhomogeneity in the similarity and frequency distribution of words in a language (Lindblom, 1990). To explain, differential frequency of occurrence makes likely words more recognizable than unlikely words. As well, the fewer the words that are similar to a target word, the more readily it is identified. A corollary of this premise provided the scheme for the Easy/Hard lists: enhanced phonetic acuity is more beneficial in identifying a less likely word from a dense lexical neighborhood of more likely words than in identifying a more likely word from a sparse lexical neighborhood of less likely words. If the crucial dimension promoting perceptual

tuning to a new talker is phonetic, then the effective exposure conditions should be identifiable by the enhanced recognition of Hard words.

In the precedent for the present study, both Easy and Hard word recognition improved when the exposure sentences and the test words matched in acoustic form and in idiolect. That is, when the exposure sentences and isolated words were both sinewave speech (Remez, Rubin, Pisoni & Carrell, 1981; Remez, 2008), and the individual whose natural utterances were used to derive the synthesis was the same in the exposure sentences and in the words, performance improved. When only the acoustic form matched (sine-wave sentences produced by a different talker) or idiolect alone matched (natural sentences spoken by the same talker whose utterances were used to model the isolated words) performance was unaffected in comparison to a control condition of no exposure. These conditions are shown schematically in Table 1. A spectrographic comparison of a natural utterance of the word GIRL and its sine-wave replica are shown in Figure 1.

The present project builds on the prior project, in which the phonemic composition of the exposure sentences was unrestricted, and across the set of a dozen and a half 3 s utterances, the phoneme classes of English were well represented. With the understanding that exposure to both acoustic variation and idiolect mattered, the present project attempts to determine the aspects of idiolect necessary to elicit talkerspecific perceptual tuning. In a new test reported here, a variety of exposure sentences was used to create a gradient of phonemic restriction. These items differed in phoneme composition. The sentence set including the phone class of *voiced obstruents* was uniquely effective in promoting intelligibility gain.

2. Method

2.1. Acoustic test materials

Two sets of test items were used in a procedure to estimate the phonemic causes of the enhancement of intelligibility observed when a listener becomes familiar with the speech of a new talker. The first set included 7 types of 10 sentences each; these were used as exposure items. Each type of sentence differed in the phonemic restriction applied to its linguistic composition. Type 1 was composed solely of vowels and liquid consonants (example: I worry while you are away); type 2 was composed of vowels and liquid and nasal consonants (example: I am well known among men); type 3 was composed of vowels, liquids and voiced fricative consonants (example: The weather is usually lovely); type 4 was composed of vowels, liquids, nasals and voiceless fricative consonants (example: All mice seem similar from far away); type 5 was composed of vowels, liquids, nasals and voiced stop consonants (example: Blend your red and blue dye); type 6 was composed of vowels, liquids, nasals and voiceless stop consonants (example: In winter I park my car near town); type 7 was phonemically unrestricted (example: The beauty of the view stunned the young boy). Types 1 through 6 had been prepared and benchmarked for intelligibility in a project on the proficiency of sine-wave synthesis (Remez, Dubowski, Davids, Thomas, Paddu, Grossman & Moskalenko, 2011); these measures revealed roughly equivalent intelligibility across the set of sentences. Type 7 was drawn from the sentence set of Experiment 2 of Remez et al., (2011a). Overall, the provenance of the sentences was diverse (Huggins & Nickerson, 1985;

Kalikow, Stevens & Elliott, 1977; Remez et al., 2011; Stubbs & Summerfield, 1990; IEEE, 1969). A complete sentence list appears in Appendix A.

The second set of test materials were the Easy/Hard word lists, which consisted of 74 nominally Easy words and 74 nominally Hard words, conforming to the designation described in section §1.2. of the *Introduction*. An Easy word was a monosyllabic English word selected from a sparse lexical neighborhood in which it was the highest frequency item amid words of lower frequency of occurrence. A Hard word was the complement: a monosyllabic word of low incidence with many lexical neighbors each of which was more frequent. The Easy/Hard target words differed, therefore, in three characteristics: mean frequency of occurrence (310 vs. 12 instances per million words, according to the norms of Kucera & Francis, 1967), mean neighborhood density (14 vs. 27 neighbors, estimated using the technique of Luce & Pisoni, 1998), and the mean frequency of occurrence of the lexical neighbors (38 vs. 282 occurrences per million). Despite this variation, the words in both sets had been reported independently as highly familiar, with an average of 6.25 on a 7-point familiarity scale (Nusbaum, Pisoni, & Davis, 1984). Additional descriptive details appear in Bradlow and Pisoni (1999); a list of the words appears in Appendix B.

The specific synthetic utterances used in this study were prepared by Remez et al. (2011a, 2011b). All of the natural utterances had been spoken by the same talker, author R. E. R. The sentences and words had been recorded direct to disk at a sampling rate of 44.1kHz with 16 bit amplitude resolution. The natural utterances were edited digitally, and the spectra were analyzed by hand to produce parameters for a sine-wave synthesizer. The synthesizer converted the frequency and amplitude estimates of three vocalic formants, the intermittent nasal murmurs and fricative formants, and the brief bursts and aperiodic transients into a set of time-varying sinusoids stored in sampled data format. Because the resulting sine-wave items were faithful to the spectrotemporal pattern of the original natural utterances, they were effective in evoking impressions of the phonetic properties that constitute dialect and idiolect, among other personal properties; tests show that sine-wave patterns derived from natural spectra are sufficient to identify the talker who spoke the original natural utterance, notwithstanding the absence of natural vocal quality (Remez et al., 1997). For use in listening tests, the items were transferred losslessly to compact disc. At the time of testing, the nominal level of 68 dB SPL was set and the items were delivered to listeners seated in a sound-attenuating chamber via Beyerdynamic DT770 headphones.

2.2. Procedure

There were seven test conditions, each with the same format. An exposure interval occurred first, consisting of a presentation of sentences in a transcription task. Following the exposure interval, a test of open set spoken word recognition occurred in which the Easy/Hard words were presented for identification. In an exposure test, 10 sentences were presented, all of a type. Each sentence was presented 5 times with a 1 s pause between recurrences and a 3 s pause between sentences. A participant was instructed to transcribe the sentence in a specially prepared booklet. However, the first five sentences in the exposure phase were printed in the test booklet when the procedure began, and the participants were asked simply to listen to those items while following the transcription provided for them. We asked them

to begin transcribing when the sixth sentence occurred. A brief intermission occurred after the exposure test, and then the Easy/Hard word identification test was conducted. On every trial, a single word was presented twice for identification, separated by 1 s of silence. There were 3 s of silence between words and 6 s of silence at the end of every tenth item, to help the participants keep track of the procession of test items. A listener was instructed to write each word on the appropriate place in the test booklet.

2.3. Participants

Ninety-one volunteers from the undergraduate population of Barnard College and Columbia University participated in the test; seven participants were dismissed for failing to follow instructions or withdrew during the test runs, leaving 84 who contributed to the measures. Each was randomly assigned to one of the seven exposure conditions. Every one self-identified as a native speaker of English, and at the time of testing disclosed a clinical history free of articulation disorder, hearing impairment or communicative difficulty. Listeners were naïve with respect to sine-wave speech, none had participated in an experiment using such acoustic items, and no participant was familiar formally or informally with the natural speech of the talker whose utterances were the models for the sine-wave items used in this procedure.

3. Results

3.1. General features of the analysis

Each subject contributed 148 scores to the analysis, the incidence of Easy words and Hard words that were correctly identified in the tests of the effectiveness of exposure. Measures from 12 participants who completed the Null exposure condition reported in Remez et al. (2011) were added to the data from the seven exposure conditions in the current experiment. Each trial in a word identification test was coded as incorrect (0) or correct (1). Descriptive statistics for the effects of word type and exposure levels reflect the proportion of each category of words that a listener identified. Figure 2 presents a plot of the group performance in each of the paired tests of spoken word identification; for comparison, the performance on the Easy/Hard word recognition tests in the Null exposure condition of Remez et al. (2011a) is also plotted.

Logistic/binomial mixed-effects regression analyses examined whether the fixed effects of word type (Easy or Hard) and exposure condition (8 levels) influenced the likelihood of correct word identification in the listening tests. All analyses included random intercepts for subjects and for words, with random slopes for word type over subjects and random slopes for exposure condition over words. The fixed effect of Word Type was contrast coded with Easy as the reference level (−0.5, 0.5).

3.2. Models with Easy and Hard words together

One model of the measures assessed whether performance differed across Easy and Hard words, and whether word identification following each of the sine-wave sentence exposure conditions differed from the Null exposure condition. This analysis used treatment coding with the Null exposure condition as the baseline level for comparison with each of the seven

exposure conditions. The fixed effects from an additive model appear in Table 2. The model revealed that the effect of word type was significant; Hard words yielded fewer correct responses than Easy words overall [(0.31 < 0.47); $\beta = -1.09$ (0.30), $Z = -3.65$, $p = 0.0003$]. With respect to the exposure conditions, only the condition that included Voiced Obstruents (VLNC+) yielded a significantly greater likelihood of correct word identification than the Null exposure condition (0.50 > 0.33); $\beta = 0.96$ (0.42), $Z = 2.30$, $p = 0.02$. The Unrestricted condition showed a similar trend (0.44 > 0.33), but the comparison was marginally significant in the analysis; $\beta = 0.75$ (0.41), $Z = 1.83$, $p = 0.07$. None of the other levels of exposure condition differed from the Null exposure condition.

A second model compared the exposure condition that included Voiced Obstruents (VLNC+V) with the other exposure conditions. In this case, treatment coding for the exposure conditions set the VLNC+V condition as the baseline level for comparison with the other seven exposure conditions. The fixed effects from an additive model appear in Table 3. The results of this model reveal that most of the exposure conditions differed from the VLNC+V condition, with the exception of the Unrestricted and VL conditions, and with a marginal difference from the VLNC-V condition (0.50 vs. 0.44); $\beta = -0.72$ (0.41), $Z = -1.77$, $p = 0.08$.

3.3. Models with Easy and Hard words separated

Parallel models on Easy and Hard word performance were performed separately, and these results basically echo the analyses including both measures. A pair of models first assessed the effect of exposure condition on performance of Easy and Hard words separately, using the Null exposure condition as the baseline for comparison with the other conditions. As shown in Tables 4a and 4b, the only conditions that differed from the null condition were the condition with Voiced Obstruents and the Unrestricted conditions (marginal).

A second pair of models assessed the effect of exposure condition on performance of Easy and Hard words separately, using the VLNC+V condition as the baseline for comparison with the other conditions. As shown in Tables 4a and b, all conditions except the Unrestricted and VL conditions differed from the VLNC+V condition for performance on easy words. A similar pattern was observed in the case of Hard words, except that differences for the VLF-V and VL conditions were marginal.

It is surprising that the benefit of exposure proved to be so limited across the conditions of exposure, specifically, to sentences featuring voiced obstruent consonants. To gauge the likelihood that this singular effect should be attributed to a close match in segmental inventory between the exposure sentences and test words, we applied a backof-the-envelope method to tally the phonemic matches between exposure sentences and test words. Table 5 shows the word counts (out of 148) conforming to the phoneme class restrictions that applied in the composition of the sentence types. From the pattern of this little exercise, the same intelligibility gain would be expected of three exposure conditions, based on a straightforward premise that an improvement in recognition performance for a word in the recognition test should follow from the presence of its phonemic constituents in the exposure sentences. Instead, it appears as though the efficacy of exposure differs inherently

across the phoneme classes in promoting a gain in intelligibility when a listener becomes acquainted with an unfamiliar talker.

4. Discussion

An introduction to a new talker imposes a perceptual challenge for a listener. First, the listener must become acquainted with the sound of the talker's voice, specifically, the acoustic manifestations of vocalization that are used for potential phonetic and personal distinctions. Second, the listener must apprehend the range of phonetic expression of the phoneme contrasts, which incorporate properties of dialect and idiolect. Although the individuals who share a language use a common inventory of words to compose expressions, the details of any utterance are bound to convey the phonetic properties of each talker's linguistic community, personal articulatory habits, and speaking style. The perceptual tuning that ensues on the part of the listener might be partial and gradual (Nygaard, et al., 1994; Sheffert, Pisoni, Fellowes & Remez, 2002), although measures reveal that this cognitive function can also be fast and long lasting (Clarke & Garrett, 2004; Eisner & McQueen, 2006). It also appears to be tagged to a specific talker, rather than recalibrating a perceiver's general perceptual standards for acoustic-to-phonetic projection (cf. Khalighinejad, Cruzato da Silva & Mesgarani, 2017). But, why would voiced obstruent consonants be especially salient or effective in short- or long-term tuning to a talker's habits?

Following Jakobson, voiced obstruents belong to the phonemic class of *marked* consonants (Jakobson, 1972; see, also, Hume 2011), which oppositions within a hierarchy are distinguished by the presence of an attribute in comparison to a neutral, or unmarked, instance. Much as a sonorant and open vowel nucleus of a syllable is distinguished from a closed and voiceless consonant at its onset, the presence of voicing in a consonant marks that segment, in contrast to the neutral form, which is voiceless. If an aspect of perceptual tuning to an unfamiliar talker obliges a listener to note the relation between neutral and marked phonetic forms, then voiced obstruents might have extra salience when a new talker's speech is encountered. As a working hypothesis, this speculation offers a clear conjecture to investigate. Alternatively, it will be useful to determine whether a deflationary conclusion is warranted instead, in which no general principle of contrast hierarchy applies. In a reduced sort of explanation, for instance, the present finding might indicate no more than attention to an idiosyncratic phonetic habit of a single talker's expression. New tests with new model talkers will be useful in resolving this empirical question.

Nonetheless, it is unlikely that sine-wave speech uniquely promotes the salience of voiced obstruent consonants among English phonemic types and phonetic variations. Because the technique of sine-wave synthesis eliminates the natural acoustic products of vocalization, sine-wave sentences lack the pulsing structure of glottal excitation which, in natural speech, is responsible for the sound of vocal quality. In consequence, neither broadband resonances nor harmonic spectra are present, conferring a distinctly nonvocal quality to a sine-wave utterance. Moreover, an intelligible sine-wave pattern lacks a component following the fundamental frequency of phonation. In fact, a tone set to the phonatory frequency fails to cohere, perceptually, with the tones replicating the vocal resonances (Remez & Rubin, 1984). Perceptual tests with sine-wave sentences which targeted the projection of acoustic to

segmental and suprasegmental attributes have shown that listeners actually form an impression of vocal pitch, however implausible, from the frequency excursions of the sinusoid replicating the lowest frequency formant (Remez & Rubin, 1984; 1993), while also using that tone as the basis for segmental phonetic impressions. Accordingly, the perception of consonant voicing in a sine-wave utterance derives from the effects of the voicing contrast on the supralaryngeal resonances (Liberman, Delattre & Cooper, 1952). Without an explicit acoustic correlate of phonation, the resonance pattern formed by the composite of individual acoustic constituents of natural speech — the whistles, hisses, clicks, buzzes and hums — is preserved in the sinusoidal replica, and this aggregation is causally effective for phonetic perception, even in a tonal language (Feng, Xu, Zhou, Yang & Yin, 2012). If a listener's experience of a sine-wave sentence lacks the typical procession of vocal qualia, the tolerance of this deficit co-occurs with an impression of an intelligible message spoken by an identifiable talker. Sine-wave consonants differing in voicing are readily transcribed by listeners without relying on semantic, syntactic or lexical context (Remez, 2008), and a challenge posed by the present set of results is to understand how the voiced subset of obstruent consonants provided an unanticipated benefit, perceptually, as perceivers became accustomed to this individual talker.

Theoretically, the present finding is not quite alignable with the alternative conceptualizations of talker identification proposed by Kreiman & Sidtis (2011). Their magisterial review concerned the psychoacoustic and perceptual literature of voice identification, individual talker learning and talker recognition. In portraying these rangy topics, two models were offered to characterize the opportunities and outcomes. One alternative views a talker as a complex pattern of attributes, and gaining familiarity with an individual's speech therefore entails a commitment to notice and to memorize a diverse assortment of idiosyncrasies. For this hypothetical class of talkers, we must presume that the capability of transfer from exposure samples to novel utterances of the kind observed here warrants perceptual resolution of many and diverse attributes. From the precedent project of Remez et al., (2011a), we know that a listener experiences a boost in intelligibility only if the exposure conditions include an opportunity to sample the acoustic properties that constitute a sine-wave talker's unusual speech spectrum. To describe this in the manner of Kreiman & Sidtis, the perceiver becomes aware of the variation in unfamiliar acoustic dimensions that correspond to known phonetic gradients or phonemic types.

In contrast, the key premise of a second alternative model proposed by Kreiman & Sidtis is that some talkers produce a clear feature which is singularly distinctive, and which can serve as a reliable personal index in the speech of such individuals. Extrapolating from that description, we might expect the sine-wave talker of this study to exhibit a single acoustic or phonetic feature by which to be identified. Could the voicing feature of obstruent consonants play this role? The benefit to intelligibility of noticing this sort of attribute is uncertain, because of its limited distribution. Moreover, the class of voiced obstruents appeared to have greater value than other phone classes of equivalent potential for producing a boost in intelligibility (see Table 2). In this regard, the hypothetical alternatives provided in Kreiman & Sidtis do not quite match the pattern of results here.

Overall, we found that the sine-wave talker of the current study produced utterances including voiced obstruent consonants in a way that permitted perceivers to induce a robust characterization of the talker's phonetic repertoire. Perhaps this was attributable to the centrality of the binary voicing contrast in English. It is simply not known whether the salience of voicing in talker learning would typify languages with 3- or 4-way voicing contrasts; or whether other features of phonemic contrast are available for this dual function, one to distinguish words and the other to distinguish talkers.

Acknowledgment

The authors are grateful to Stavroula Koinis, Natalie Porter and Nina Paddu for advice and encouragement in developing this project. This research was supported by a grant from the National Institute on Deafness and Other Communication Disorders to author R. E. R. (DC000308).

Appendix A. Sentences of Varying Phonemic Restriction

1. Vowels and liquid consonants

I owe you a yoyo.
 A war ally will rule Iowa.
 Where were you well?
 Larry wore a laurel a year early.
 Lower your arrow Ella.
 Why are you weary?¹
 Will your lawyer allow our error?
 I worry while you are away.
 Are you aware I will roll away?
 We all wear a rare yellow wool.

2. Vowels, liquids and nasal consonants

I am well known among men.¹
 I normally iron all morning.
 A royal memorial will remain.³
 Are you a loyal union man?³
 I owe no one any money.¹
 You lie in an alarming manner.¹
 I will marry you in May.³
 I am wearing my maroon one.¹

¹.Huggins & Nickerson (1985).

³.Stubbs & Summerfield (1990).

When will our yellow lion roar?¹

You were wrong all along.¹

3. Vowels, liquids, and voiced fricatives

Will they allow you a lawyer?

Lower the level of the revolver.

The weather is usually lovely.

Our rival will arrive early.

Is Lily loyal or evil?

Reserve the olives of the rural villa.

While I weigh the leather you use the razor.

Liver is always vile.

Our rival loves the zoo.

Will you reveal the loser of the war?

4. Vowels, liquids, nasals, and voiceless fricative consonants

He swore he fell on a shoe in your house.

All mice seem similar from far away.

She will sell a flower near a sea shore.

A rainy Fall will follow a sunny Summer.

Will you seriously follow her?

How will you see a show for free?

If you wear a high heel shoe, you will suffer.

If you sin, you will learn your lesson.

A frail woman will feel safe on her sofa.

Soon our chef will weigh some flour.

5. Vowels, liquids, nasals, and voiced stop consonants

An aid will guide you around our building.

A big wall would be a good border.

Bobby did a good deed.¹

Bend a band aid around your bloody elbow.

I needed a brand new rubber band.

Are you bored by your own name?

Do you abide by your bid?¹

A greedy boy died.¹

A deer and a bear will gladly dawdle in your garden.

Blend your red and blue dye.

6. Vowels, liquids, nasals, and voiceless stop consonants

A parrot in a crate will talk all night.

I cannot tell a tale too well.

Our turtle ate a tiny kiwi.

In winter I park my car near town.

Can we keep your kite until tomorrow?

I can knit one mitten per minute.

Take a copy to Pete.¹

Tell my uncle not to take our apple pie.

We met you on time at an airport terminal.

Can you write a term paper in a week?

7. Unrestricted phoneme composition

The beauty of the view stunned the young boy.⁴

The steady drip is worse than a drenching rain.⁴

The bark of the pine tree was shiny and dark.⁴

The drowning man let out a yell.²

They took the axe and the saw to the forest.⁴

The boy was there when the sun rose.⁴

Her purse was full of useless trash.⁴

The sandal has a broken strap.²

Two blue fish swam in the tank.⁴

A pencil with black lead writes best.⁴

¹.Huggins & Nickerson (1985).

².Kalikow, Stevens, & Elliott (1977).

³.Stubbs & Summerfield (1990).

⁴.IEEE (1969).

⁴.IEEE (1969).

².Kalikow, Stevens, & Elliott (1977).

Appendix B. Easy and Hard Words

Easy Words				
balm	firm	king	pool	teeth
both	five	league	pull	theme
cause	food	learn	put	thick
chain	fool	leg	reach	thing
check	full	live	real	thought
chief	gas	long	roof	vice
curve	gave	lose	rough	voice
death	girl	love	serve	vote
deep	give	mouth	shall	was
dirt	god	move	shape	wash
does	hung	neck	ship	watch
dog	jack	noise	shop	wife
down	job	page	size	work
faith	join	path	soil	young
fig	judge	peace	south	
Hard Words				
ban	dame	hurl	moat	sill
bead	den	kin	mole	soak
beak	doom	kit	mum	suck
bean	dune	knob	pad	tan
bud	fade	lace	pat	teat
bug	fin	lad	pawn	toot
bum	goat	lame	pet	wad
bun	gut	lice	pup	wade
chat	hack	mace	rat	wail
cheer	hag	main	rhyme	wed
chore	hash	mall	rim	weed
cod	hick	mat	rout	white
comb	hid	mid	rum	whore
con	hoot	mitt	rut	wick
cot	hum	moan	sane	

References

- Abercrombie D (1967). *Elements of General Phonetics*. Chicago: Aldine.
- Allen JS, & Miller JL (2004). Listener sensitivity to individual talker differences in voice onset-time. *Journal of the Acoustical Society of America*, 115, 3171–3183. DOI: 10.1121/1.1701898 [PubMed: 15237841]
- Bradlow AR, Pisoni DB (1999). Recognition of spoken words by native and nonnative listeners: Talker-, listener-, and item-related factors. *Journal of the Acoustical Society of America*, 106 2074–2085. DOI: 10.1121/1.427952 [PubMed: 10530030]
- Bricker PD, & Pruzansky S (1976). Speaker recognition In Lass NJ (Ed.), *Contemporary Issues in Experimental Phonetics* (pp. 295–326). New York: Academic Press.

- Clarke CM, & Garrett MF (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, 116, 3647–3658. DOI: 10.1121/1.1815131 [PubMed: 15658715]
- Eisner F, & McQueen JM (2006). Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America*, 119, 1950–1953. DOI: 10.1121/1.2178721 [PubMed: 16642808]
- Feng Y-M, Xu L, Zhou N, Yang G, & Yin S-K (2012). Sine-wave speech recognition in a tonal language. *Journal of the Acoustical Society of America*, 131, EL133–EL138. DOI: 10.1121/1.3670594 [PubMed: 22352612]
- Goldinger SD, Pisoni DB, & Logan JS (1991). On the nature of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 152–162. DOI: 10.1037/0278-7393.17.1.152
- Huggins AWF, and Nickerson RS (1985). Speech quality evaluation using phoneme-specific sentences. *Journal of the Acoustical Society of America*, 77, 1896–1906. DOI: 10.1121/1.391941 [PubMed: 3998299]
- Hume E (2011). Markedness In van Oostendorp M, Ewen C, Hume E & Rice K (eds.), *Companion to Phonology, Vol. 1: General Issues and Segmental Phonology* (pp. 79–106). Oxford: Blackwell.
- IEEE (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, AU-17, 225–246. DOI: 10.1109/IEEESTD.1969.7405210
- Jakobson R (1972). Verbal communication. *Scientific American*, 227:3, 72–80. DOI: 10.1038/scientificamerican0972-72 [PubMed: 4628768]
- Johnson K, & Mullenix JW (1997). *Talker Variability in Speech Processing*. New York Academic Press.
- Kalikow DN, Stevens KN, and Elliot LL (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, 61, 1337–1351. DOI: 10.1121/1.381436 [PubMed: 881487]
- Khalighinejad B, Cruzato da Silva G, & Mesgarani N (2017). Dynamic encoding of acoustic features in neural responses to continuous speech. *Journal of Neuroscience*, 37, 2176–2185. DOI:10.1523/JNEUROSCI.2383-16.2017 [PubMed: 28119400]
- Kraljic T, Brennan SE, & Samuel AG (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, 107, 54–81. DOI: 10.1111/j.1467-9280.2008.02090.x [PubMed: 17803986]
- Kreiman J, & Sidtis D (2011). *Foundations of Voice Studies*. Oxford: Wiley-Blackwell.
- Kucera H, & Francis WN (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.
- Ladefoged P (1967). *Three Areas of Experimental Phonetics*. London: Oxford University Press.
- Lieberman AM, Cooper FS, Shankweiler DP, & Studdert-Kennedy M (1967). Perception of the speech code. *Psychological Review*, 74, 431–461. DOI: 10.1037/h0020279 [PubMed: 4170865]
- Lieberman AM, Delattre PC, & Cooper FS (1952). The role of selected stimulusvariables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 65, 497–516. DOI: 10.2307/1418032 [PubMed: 12996688]
- Lindblom B (1990). Explaining phonetic variation: A sketch of the H & H theory In Hardcastle WJ and Marchal A (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Dordrecht: Kluwer.
- Luce PA, & Pisoni DB (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19, 1–38. DOI: 10.1097/00003446-199802000-00001 [PubMed: 9504270]
- Mullennix JW, & Pisoni DB (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, 47, 379–390. DOI: 10.3758/BF03210878 [PubMed: 2345691]
- Nolan F (1983). *The Phonetic Basis of Speaker Recognition*. Cambridge: Cambridge University Press.
- Nusbaum HC, Pisoni DB, & Davis CK (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words *Research in Speech Perception*, Progress Report 10 (pp. 357–376). Bloomington Indiana: Indiana University.
- Nygaard LC, Sommers MS, & Pisoni DB (1994). Speech perception as a talkercontingent process. *Psychological Science*, 5, 42–46. DOI: 10.1111/j.1467-9280.1994.tb00612.x [PubMed: 21526138]

- Remez RE (2008). Sine-wave speech In Izhikovitch EM (Ed.), *Encyclopedia of Computational Neuroscience* (pp. 2394). (Cited as Scholarpedia, 3, 2394) DOI: 10.4249/scholarpedia.2394
- Remez RE (2010). Spoken expression of individual identity and the listener In Morsella E (Ed.), *Expressing Oneself/Expressing One's Self: Communication, Cognition, Language, and Identity* (pp. 167–181). New York: Psychology Press.
- Remez RE, Dubowski KR, Broder RS, Davids ML, Grossman YS, Moskalenko M, Pardo JS, & Hasbun SM (2011a). 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 [PubMed: 21443384]
- Remez RE, Dubowski KR, Davids ML, Thomas EF, Paddu NU, Grossman YS, & Moskalenko M (2011b). Estimating speech spectra by algorithm and by hand for synthesis from natural models. *Journal of the Acoustical Society of America*, 130, 2173–2178. DOI: 10.1121/1.3631667 [PubMed: 21973371]
- Remez RE, Fellowes JM, & Rubin PE (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 651–666. DOI: 10.1037//0096-1523.23.3.651 [PubMed: 9180039]
- Remez RE, & Rubin PE (1984). Perception of intonation in sinusoidal sentences. *Perception & Psychophysics*, 35, 429–440. DOI: 10.3758/BF03203919 [PubMed: 6462869]
- Remez RE, & Rubin PE (1993). On the intonation of sinusoidal sentences: Contour and pitch height. *Journal of the Acoustical Society of America*, 94, 1983–1988. DOI: 10.1121/1.407501 [PubMed: 8227742]
- Remez RE, Rubin PE, Pisoni DB, & Carrell TD (1981). Speech perception without traditional speech cues. *Science*, 212, 947–950. DOI: 10.1126/science.7233191 [PubMed: 7233191]
- Remez RE, & Thomas EF (2013) Early recognition of speech. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4, 213–223. DOI: 10.1002/wcs.1213. [PubMed: 23926454]
- Sheffert SM, Pisoni DB, Fellowes JM & Remez RE (2002). Learning to recognize talkers from natural, sine-wave and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 1447–1469. DOI: 10.1037//0096-1523.28.6.1447 [PubMed: 12542137]
- Stubbs RJ, & Summerfield Q (1990). Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 87, 359–372. DOI: 10.1121/1.399257 [PubMed: 2299046]

Figure 1.

Spectrographic representation of the easy word GIRL. A token of natural speech used as the model for the sine-wave version is shown in the left of the figure; the three-tone sine-wave replica is shown in the right of the figure. Note the absence of aperiodic release burst, glottal pulsing and broadband resonances in the sine-wave version, which replicates the estimated frequency and amplitude variation of the natural spectrotemporal pattern, yet does so in three time-varying sinusoids. Sine-wave replicas of natural utterances express the aggregate pattern of acoustic effects of natural vocalization, yet lack the momentary acoustic components (the “speech cues”) characteristic of natural speech (after Remez et al., 2011a)

Figure 2.

The mean identification performance of the seven tests of the effects of exposure on the identification of Easy and Hard words. Along the *x*-axis, the exposure conditions are arrayed, described by the feature composition of the sentence set. Note that an eighth condition, Null Exposure reported in Remez et al. (2011a), is also included as a standard for comparison. In each exposure condition, the identification performance of Easy and Hard words, expressed as a proportion of possible performance, is shown in a pair of bars. The filled bars show Easy word performance, the open bars show the Hard word performance. Error bars portray the standard error of each group.

Table 1.

Exposure-driven Enhancement of Spoken Word Identification (Remez et al., 2011a).

EXPOSURE	MATCHING DIMENSION	TEST OF EXPOSURE	BENEFIT OF EXPOSURE
No Exposure	none	Easy/Hard Words SW_{RER}	$\emptyset$
Sentences SW_{RER}	acoustic form, idiolect	Easy/Hard Words SW_{RER}	+
Sentences $Natural_{\text{RER}}$	idiolect	Easy/Hard Words SW_{RER}	$\emptyset$
Sentences SW_{JSP}	acoustic form	Easy/Hard Words SW_{RER}	$\emptyset$

A description of the conditions used in a study of exposure-based gain in intelligibility. The initials of the talker appears in subscripts; the acoustic form was either sampled natural speech or sine-wave speech (SW) derived from the natural models. A benefit of exposure, observed as an increase in performance level in word identification, is marked with a plus sign (+), and was observed only in the condition in which both acoustic form and idiolect matched between exposure sentences and test words

Table 2.

Analysis of Easy Hard Word Identification Performance Null Exposure condition as base, model converged

	<i>Estimate</i>	<i>SE</i>	<i>Z</i>	<i>p</i>
Intercept	-1.10	0.34	-3.28	0.0010**
Easy vs. Hard	-1.09	0.30	-3.65	0.0003***
<i>Comparisons with No Exposure Condition</i>				
VL	0.58	0.41	1.42	0.157
VLN	-0.14	0.41	-0.35	0.725
VLFF+V	0.13	0.41	0.32	0.749
VLFF-V	-0.13	0.42	-0.30	0.766
VLNC-V	0.24	0.41	0.58	0.560
VLNC+V	0.96	0.42	2.30	0.021*
Unrestricted	0.75	0.41	1.83	0.067

Results of a logistic/binomial mixed-effects regression analysis of the fixed effects of exposure condition (8 levels) on identification of word type (Easy or Hard). Here, the effects relative to Null Exposure are shown. The fixed effect of Word Type was contrast coded with Easy as the reference level (-0.5, 0.5). The model revealed that the effect of word type was significant; identification performance differs for Hard words and Easy words overall. Only the condition that included Voiced Obstruents (VLNC+V) differed from Null exposure condition.

Table 3.

Analysis of Easy Hard Word Identification Condition with Voiced Obstruents (VLNC+V) as base, model converged

	<i>Estimate</i>	<i>SE</i>	<i>Z</i>	<i>p</i>
Intercept	−0.15	0.32	−0.45	0.6524
<i>Easy vs. Hard</i>	−1.09	0.30	−3.65	0.0003***
<i>Comparisons with VLNC+V Condition</i>				
Unrestricted	−0.21	0.40	−0.52	0.603
VLNC-V	−0.72	0.41	−1.77	0.077
VLF-V	−1.08	0.41	−2.66	0.008**
VLF+V	−0.83	0.41	−2.03	0.043*
VLN	−1.10	0.41	−2.71	0.007**
VL	−0.38	0.41	−0.92	0.357
none	−0.96	0.42	−2.30	0.021*

Results of a logistic/binomial mixed-effects regression analysis of the fixed effects of exposure condition (8 levels) on identification of word type (Easy or Hard). Here, the effects relative to the exposure conditions containing Voiced Obstruent consonants (VLNC+V) are shown. Again, the fixed effect of Word Type was contrast coded with Easy as the reference level (−0.5, 0.5). Most of the exposure conditions differed from the VLNC+V condition, with the exception of the Unrestricted and VL conditions.

Table 4.

Analysis of Word Identification Performed Separately on Easy and Hard Word List

<i>4a: Easy Words</i>				
	<i>Estimate</i>	<i>SE</i>	<i>Z</i>	<i>p</i>
(Intercept)	−0.54	0.36	−1.50	0.133
<i>Comparisons with Null Exposure Condition</i>				
VL	0.537	0.40	1.34	0.179
VLN	−0.124	0.40	−0.31	0.756
VLF+V	0.157	0.40	0.39	0.694
VLF-V	0.003	0.40	0.01	0.994
VLNC-V	0.185	0.40	0.46	0.643
VLNC+V	1.060	0.40	2.65	0.008**
Unrestricted	0.769	0.40	1.93	0.054
<i>4b: Hard Words</i>				
	<i>Estimate</i>	<i>SE</i>	<i>Z</i>	<i>p</i>
(Intercept)	−1.83	0.40	−4.62	0.000***
<i>Comparisons with Null Exposure Condition</i>				
VL	0.551	0.49	1.13	0.261
VLN	−0.002	0.49	−0.01	0.996
VLF+V	0.366	0.49	0.75	0.455
VLF-V	0.562	0.49	1.14	0.253
VLNC-V	0.151	0.49	0.31	0.759
VLNC+V	1.375	0.49	2.82	0.005**
Unrestricted	0.913	0.49	1.87	0.062

The effect of exposure condition on identification performance with Easy (Table 4a) and Hard (Table 4b) words, assessed separately, using a logistic/binomial mixed-effects regression analysis with the Null exposure condition as the baseline for comparison with the other conditions. The only exposure condition that differed from the Null condition was the condition including Voiced Obstruents; and the Unrestricted conditions differed marginally.

Table 5.

Analysis of Word Identification Performed Separately on Easy and Hard Word Identification Performance

<i>5a: Easy Words</i>				
	<i>Estimate</i>	<i>SE</i>	<i>Z</i>	<i>p</i>
(Intercept)	0.52	0.36	1.44	0.151
<i>Comparisons with VLNC+V Condition</i>				
unrestricted	−0.29	0.40	−0.73	0.465
VLNC-V	−0.87	0.40	−2.19	0.028*
VLF-V	−1.06	0.40	−2.65	0.008**
VLF+V	−0.90	0.40	−2.26	0.024*
VLN	−1.18	0.40	−2.97	0.003**
VL	−0.52	0.40	−1.31	0.189
none	−1.06	0.40	−2.66	0.008**
<i>5b: Hard Words</i>				
	<i>Estimate</i>	<i>SE</i>	<i>Z</i>	<i>p</i>
(Intercept)	−0.46	0.39	−1.17	0.242
<i>Comparisons with VLNC+V Condition</i>				
unrestricted	−0.46	0.49	−0.95	0.341
VLNC-V	−1.22	0.49	−2.51	0.012*
VLF-V	−0.81	0.49	−1.67	0.096
VLF+V	−1.01	0.49	−2.07	0.038*
VLN	−1.38	0.49	−2.81	0.005**
VL	−0.82	0.49	−1.70	0.090
none	−1.38	0.49	−2.81	0.005**

The effect of exposure condition on identification performance with Easy (Table 5a) and Hard (Table 5b) words, assessed separately, using a logistic/binomial mixed-effects regression analysis with the VLNC+V condition as the baseline for comparison with the other conditions. As shown in Tables 5a and 5b, all conditions except the Unrestricted and VL conditions differed from the VLNC+V condition for performance on easy words. A similar pattern was observed in the case of Hard words, except that differences for the VLF-V and VL conditions were marginal.

Table 6.**Phoneme Composition of the Easy/Hard Word Lists**

tally	Phoneme Class
3	vowels, liquids
14	vowels, liquids, nasals
5	vowels, liquids, voiced fricatives
23	vowels, liquids, nasals, voiceless fricatives
26	vowels, liquids, nasals, voiced obstruents
25	vowels, liquids, nasals, voiceless obstruents
148	unrestricted

A back-of-the-envelope method of predicting intelligibility gain from shared segmental properties. The right column shows the feature restriction exhibited of the sentences that were used in each exposure condition. The left column shows the tally of words (out of 148) in the Easy/Hard word lists exhibiting the features of the exposure sentences. A prediction by the incidence of features identifies three conditions with approximately equal benefit: vowels, liquids, nasals, voiceless fricatives; vowels, liquids, nasals, voiced obstruents; and, vowels, liquids, nasals, voiceless obstruents. However, facilitation was only observed in one of these conditions: vowels, liquids, nasals, voiced obstruents. (See *Discussion*.)