The goal of this review article is to reinvigorate interest in lipreading for the rehabilitation of and research on speech recognition in adults with acquired hearing loss. Lipreading is the recognition of speech by vision alone. The term *speechreading* is sometimes used interchangeably with *lipreading*. But *speechreading* more often refers to audiovisual (AV) speech recognition,\(^1\) so we use the term *lipreading* when referring to visual-only (VO) speech perception and recognition.

In most of the 20th century, visual speech was considered to be critical for successful face-to-face communication by individuals with hearing loss. For example, Montgomery and colleagues (Montgomery et al., 1984) observed, “It is obvious that the hard-of-hearing patient...
who hopes to maximize his or her speech reception ability must combine speechreading [lipreading] with residual hearing in an efficient way. The potential benefits of integrating auditory and visual cues are especially significant for the adult with acquired, moderate high-frequency hearing loss...” (p. 30). In light of the advances that have been made in hearing aid technologies, a fair question is whether lipreading is as important today as it was in the last century. The answer is that maximizing speech recognition with a hearing loss continues to require combining visual and auditory speech information (J. G. W. Bernstein et al., 2020). In noisy listening conditions, the ability to see the talker may overcome speech perception difficulties that a hearing aid alone cannot overcome.

**Influence of Lipreading on AV Recognition in Noise**

The intelligibility of noisy and/or degraded speech has long been known to increase when the talker can be seen as well as heard (i.e., Erber, 1969; Middelweerd & Plomp, 1987; Sumby & Pollack, 1954). Fortunately, for older adults with acquired hearing loss, AV speech integration ability may even increase with age (Dias et al., 2021; Sommers et al., 2005; Tye-Murray et al., 2016; Winneke & Phillips, 2011).

Middelweerd and Plomp (1987) estimated that the speech reception threshold for correct syllables in sentences was approximately 4 dB signal-to-noise ratio (SNR) lower with AV versus AO speech in a group of unaided older adults with thresholds ≤ 40 dB HL. In MacLeod and Summerfield (1986), AV benefit (the difference in dB speech reception thresholds for AO vs. AV speech in white noise) was considered to be a measure of lipreading ability for adults with normal hearing for words in sentences, because AV and VO scores were so highly correlated (n = 20, r = .86, p < .012). Summerfield (1992) estimated that lipreading skill enables an individual to tolerate 4-6 dB poorer SNR in speech reception threshold and that every decibel of SNR gained corresponds to an estimated gain of 10%–15% improvement in speech recognition.

L. E. Bernstein et al. (2021) reported, in a study with 57 younger adults who had normal hearing, the correlation between AV and VO scores on a test of open set recognition of words in sentences (with the audio in speech-shaped noise) was Pearson r = .65 (n = 57, p < .001) before lipreading training and r = .77 (p < .001) afterwards. The correlation between AV and auditory-only (AO) speech recognition was much smaller at both times (pretraining, r = .27; posttraining, r = .24, p < .05).

Thus, the effect size of visual speech for AV speech recognition was 42%–59% variance accounted for, whereas the effect size of AO speech was only 6%–7% variance accounted for. There are correlations reported in the literature that are smaller between VO and AV versus AO and AV speech recognition (e.g., Dias et al., 2021); nevertheless, it is clear that lipreading ability is critical in achieving the full potential benefits of combining auditory and visual speech information.

Importantly, there are very large individual differences in lipreading ability among adults who have experienced normal hearing throughout most of their lives (Auer & Bernstein, 2007; L. E. Bernstein et al., 2000; Hall et al., 2005; Mohammed et al., 2005). For example, in two studies, lipreading scores for a large group of younger adults with normal hearing averaged around 20% words correct in sentences, although the ranges were from about 0% to above 60% (Auer & Bernstein, 2007; L. E. Bernstein et al., 2000). Therefore, with most adults’ lipreading at much below ceiling, and with lipreading ability a strong factor in AV speech recognition, effective lipreading training for adults with acquired hearing loss could deliver important benefits.

Given the availability of sophisticated hearing technologies, from our perspective, the primary goal for lipreading training for adults with acquired hearing loss is not routine reliance on silent lipreading but rather more accurate AV speech recognition in noisy social settings. Unfortunately, the difficulty of achieving this goal through training has led many researchers to conclude that lipreading is probably an inborn trait (e.g., Clousser, 1977; Conrad, 1977; Massaro, 1987; J. Rönningen, 1995; Summerfield, 1991). Our position is that this conclusion was made prematurely. Recently, we have been working to develop effective training paradigms. The remainder of this review article describes the rationale behind our approach.

**Outline of the Review Article**

Below, we first describe some important background about lipreading. Then, some of the older research on lipreading is described. In the section on training, theoretical work on perceptual learning is described as the rationale for two of our training experiments that we discuss. We then briefly comment on contemporary auditory-speech training research in comparison with lipreading research. The review article concludes with some remarks about AV speech recalibration effects, AV speech training, and effects of visual speech on auditory perception of processed acoustic signals.

**Perception and Recognition of Visual Speech**

In this section, characteristics of sublexical visual speech perception are discussed followed by a discussion...
of the lexical constraints that contribute to lipreading. A common assumption about lipreading is that it depends on constructive processes such as guessing or filling in when only parts of the stimulus can be recognized due to visual phonemic ambiguity (Van Tasell & Hawkins, 1981). Higher level processes, including filling-in and guessing strategies, in addition to psycholinguistic processing at the levels of word meaning and syntax, no doubt do contribute to lipreading accuracy with connected speech stimuli. However, the focus in this section is on sublexical and lexical levels of lipreading, because they are tied most directly to the visual perceptual level of lipreading, whereas the higher processing levels such as syntax are considered to be amodal. That is, for example, processing of the syntax, meaning, and pragmatics of spoken language is later in bottom-up cortical hierarchies (Hickok & Poeppel, 2007; Hickok et al., 2018) than are speech stimulus representations (L. E. Bernstein & Liebenthal, 2014).

**Sublexical Visual Speech Perception**

Phonemic perceptual confusions are probably the most common aspect of lipreading that is described in the literature. Indeed, there are frequent phonemic perceptual confusions such as between /h/ and /d/. These confusions are often explained as evidence for so-called “visemes,” visual speech categories (Fisher, 1968; Woodward & Barber, 1960). An accepted definition of “visemes” is that, “…a difference between visemes is significant, informative, and categorical to the perceiver; a difference within a viseme class is not” (Massaro et al., 2012, p. 316; cf., Peelle & Sommers, 2015). For example, the phoneme group /b, p, m/ is often considered to be a viseme whose members cannot be discriminated. As in the /b, p, m/ example, the members of a viseme often share place of articulation, and as a consequence, lipreading is frequently characterized as primarily delivering place of articulation information (e.g., Braida, 1991; Massaro, 1987).

However, the viseme as a perceptual category is not an accurate scientific description of VO phonemic perception despite its widespread acceptance. First, there is not one fixed set of visemes. What counts as a viseme depends on factors such as vowel context, the talker, and the perceiver (e.g., Owens & Blazek, 1985). Second, visemes are the outcome of applying computational and algorithmic procedures to continuous values in phoneme identification confusion data matrices (e.g., Auer & Bernstein, 1997; Walden et al., 1981). The procedures vary across studies, but they generally involve transforming phoneme confusion matrices into similarity/dissimilarity matrices and then applying a procedure such as hierarchical clustering to group phonemes together whose perceptual responses pattern similarly (Iverson et al., 1998). The level of perceptual similarity that is used to define the viseme groupings is arbitrary in the sense that researchers choose a value to operationalize the viseme concept.

Furthermore, the notion that visemes are perceived categorically, and that lipreaders mostly perceive place of articulation has been shown to be incorrect. Lipreaders can perceive phoneme voicing and manner of articulation information (e.g., L. E. Bernstein et al., 2000; Lalonde & Werner, 2019). Lipreaders also perceive prosodic information (i.e., intonation and lexical stress; e.g., L. E. Bernstein et al., 1989; Cutler & Jesse, 2021; Lansing & McConkie, 1999).

While the study of AO sublexical categorical speech perception is grounded in results obtained with both identification and discrimination paradigms (Pisoni & Lazarus, 1974), the results on viseme perception are mostly from forced choice identification paradigms. As a consequence, a critical aspect of the characterization of categorical perception is typically missing, that is, whether discrimination is at chance within viseme categories. This omission could be because discrimination is assumed to be poor, or possibly, because the VO speech stimulus carries additional non-speech facial information, which could be used for discrimination. Thus, the stimulus preparation and the discrimination paradigm must carefully control for possible non-speech contributions to discrimination performance in order for results on visual phoneme discrimination to be considered valid.

In fact, when visual phoneme same-different discrimination was tested using both natural and synthetic consonant-vowel (CV) syllable pairs, taking into account possible non-speech artifacts in the stimuli, younger adults with normal hearing reliably discriminated between phonemes that were within visemes (p < .05; Files et al., 2015). During the Files et al. Experiment 1, participants performed speeded discrimination for 31 different pairs of consonant-vowel nonsense syllables that were predicted to be at a perceptual distance of same, near, or far. Near pairs comprised consonants within visemes. Natural within-viseme pairs were discriminated above chance (p < .05; except for /kl–/hl/), and sensitivity (d') increased and response times decreased with predicted perceptual distance (p < .05).

Using several of the same stimuli as in Files et al. (2015) in an electrophysiology experiment with the same type of participants, mismatch negativity responses for phonemes within and across visemes elicited neural change detection responses (Files et al., 2013). Only perceptually far deviant stimuli evoked the visual mismatch negativity response over the left posterior temporal cortex, whereas perceptually far and near deviants evoked the response over right posterior temporal cortex. That is, the brain detected stimulus differences that were used for phoneme categories. It also detected differences that were not being used for phoneme category distinctions. We discuss later the importance of discriminable stimulus information for perceptual learning of visual speech categories.
As mentioned earlier, contrary to the notion that VO information is primarily phoneme place of articulation, lipreaders can to some extent perceive voicing and manner of articulation information (e.g., L. E. Bernstein et al., 2000; Lalonde & Werner, 2019). Voicing is an interesting articulatory feature in that it manifests in many acoustic phonetic characteristics (Lisker & Abramson, 1964). In postvocalic consonants, the voicing distinction in English is often a function of the preceding vowel duration (Raphael, 1971), which may be visible. This vowel duration cue is also a phonetic cue to visible lexical stress (Kecting et al., 2000). Several consonants that are characterized by continua manner of articulation are highly visually discriminable to adults with normal hearing. The phonemes /w/, /l/, and /r/ can each be placed in their own viseme (Iverson et al., 1998; Mattys et al., 2002) based on conventional viseme generation methods. The phonemes /b/, /p/, and /m/ have been shown to be separable into /b/, /p/ versus /m/, a manner distinction, based on discrimination results (Scheinberg, 1988).

Thus, presuppositions to the effect that VO speech information is limited to viseme categories are false. Subvisemic VO speech information has been shown to be available when experiments use appropriate paradigms. The discriminability of subvisemic information is used in lipreading words, to which we now turn.

**Lexical Constraints**

Lipreaders can recognize words despite there being less stimulus information in visual than in clear auditory speech stimuli. The possibility that words can be recognized accurately based on incomplete visual phonetic information is entirely consistent with the ability to recognize auditory spoken words based on incomplete or degraded auditory phonetic information. By “entirely consistent,” we do not intend to imply that the neural mechanisms for visual and auditory speech processing are exactly the same, which they are not (L. E. Bernstein & Liebenthal, 2014; Nidiffer et al., 2021; Venezia et al., 2017). We mean that complete phonetic information is often not required to recognize spoken words, regardless of modality.

An accepted explanation for auditory spoken word recognition with complete or incomplete stimulus information is that it is a competitive process (Luce & Pisoni, 1998; Marslen-Wilson & Tyler, 1980; McClelland & Elman, 1986; Norris et al., 1995; Weber & Scharenborg, 2012): The stimulus word activates similar words in the mental lexicon as the stimulus word unfolds in real time. Auditory word recognition is achieved when one word outcompetes the other candidate words, that is, when its phonetic cues or phonemic categories diverge from that of other words.

While there is not nearly the extent of visual spoken word recognition research in comparison with that on auditory and printed word recognition, the constraining role of the lexicon in lipreading words has been shown experimentally. In studies that have used estimates of the visual confusability of words to compute lexical neighborhood sizes (i.e., the number of visually similar words for each word in the lexicon), deaf and hearing adults identified words in small lexical neighborhoods (with few visually similar words) reliably more accurately than words in large neighborhoods ($p < .05$; Auer, 2002; Mattys et al., 2002). These effects have been further demonstrated by other researchers (e.g., Strand, 2014; Strand & Sommers, 2011).

In another study (L. E. Bernstein, 2012), deaf and normal-hearing younger adult lipreaders were shown isolated printed words and asked to select which of the two visual spoken words matched the printed target words. The two VO spoken words had the same visemes but different consonants (with viseme order held constant). If the words were internally represented exclusively in terms of viseme categories, participants would be unable to correctly match the printed target word to the VO spoken version. However, even at the most difficult level (i.e., within phoneme groups that were perceptually closer than visemes), matches were above chance ($p < .0001$). There were deaf and hearing participants whose scores were in the range of 65%–80% correct. In order to recognize words this accurately, the participants must have perceived visual subvisemic phonetic details.

When acoustic phonetic information is ambiguous or degraded, listeners are biased to hear real words. Effects of the lexicon have been shown with acoustic speech continua. For example, the so-called Ganong effect (Ganong, 1980) occurs when an ambiguous phoneme /d/-/t/ is placed in the context of dash or tash versus dash or task, and listeners identify the stimuli as the meaningful real words not the nonsense words ($p < .05$). This bias toward a lexical response is likely also important in lipreading words. In our experience, normal-hearing adults frequently make whole-word lipreading errors (L. E. Bernstein, 2018), even though they are encouraged to give partial responses if unsure. Thus, word lipreading has characteristics of lexical constraint that appear to be shared with auditory word recognition.

**Implications of Visual Phonemic Perception for Word Recognition**

Assuming that the extent of perceptible visual phonetic detail influences visual spoken word recognition, even small increases in perceived cues should have effects...
on perception. Auer and Bernstein (1997) computationally modeled different numbers of what are referred to as phonemic equivalence classes (similar to viseme groupings but across a range of perceptual confusabilities). They then recorded a digital lexicon using symbols to represent entire classes. For example, “B” would represent /b, p, m/. They demonstrated that when all of the vowels and consonants of English were represented by only 10 classes, that is, the classes, /u, o, æ, æ/, /aʊ, aʊ/, /i, e, æ, æ/, /a, a, a, a, ə, ə/, /b, p, m/, /f, v/, /l, n, k, ʃ, t, ʃ, t, ʒ, dʒ, z, s/, most words were predicted to be visually similar to other words. When the phonemes were represented by 12 classes, that is, /u, o, æ, æ/, /aʊ, aʊ/, /i, e, æ, æ/, /a, a, a, a, ə, ə/, /b, p, m/, /f, v/, /l, n, k, ʃ, t, ʃ, t, ʒ, dʒ, z, s/, and /f, tʃ, z, dʒ/), most words were predicted to be visually distinct from other words. This modeling suggests that there may be large effects of lipreading training, even if training were to successfully increase perception of only a few visual phonetic cues.

Summary of Perception and Recognition of Visual Speech

Visual speech perception is not limited to viseme categories. Adults with normal hearing can discriminate between phonemes that are within putative viseme groups, and they can use the information to identify a target word whose visemes are the same as another word’s but comprise different phonemes. Word lipreading is constrained by the lexicon, and lipreaders are biased to recognize whole words. Small improvements in VO phoneme perception could result in meaningful improvements in lipreading and, by extension, could also result in increases in AV speech recognition in noise.

Lipreading Training

We begin this section on lipreading training with a brief review of training paradigms that were investigated in the 20th century. We then focus on the neural basis of perceptual learning as revealed by research mostly on visual perception and on auditory speech recognition. We then discuss some results from our own research.

Twentieth Century Lipreading Training

The possibility of using training to improve VO and AV speech recognition was entertained by several researchers in the 20th century (e.g., DeFilippo, 1988; Gesi et al., 1992; Jeffers & Barley, 1971; Walden et al., 1981). The rationale for the training of that time was tied theoretically more to an understanding of the structure of language (i.e., phonemes, words, and sentences) than to knowledge about the effects of the tasks applied during training. The influence of the training task on speech perceptual learning has been substantially researched primarily in the 21st century.

Analytic Versus Synthetic Training

Twentieth century training paradigms were categorized as analytic or synthetic. Analytic usually referred to phoneme category training with isolated nonsense syllables, and synthetic referred to training with connected speech, not only with isolated sentences or phrases (Lesner et al., 1987; Walden et al., 1981) but also with connected texts (DeFilippo, 1988). Analytic training was thought to be potentially the more efficient and effective approach, because there are a relatively small number of phonemes in a language that give access to all of its words. In contrast, synthetic training was thought to use more ecologically valid stimuli and to engage listener strategies such as guessing and using semantic context.

During typical analytic training, participants carried out forced choice phoneme identification with nonsense syllables and received feedback that revealed the correct response. However, research with visual and/or auditory stimuli did not result in much evidence for the efficacy of the approach for adults with hearing loss (Blamey & Alcantara, 1994; Lesner et al., 1987; Montgomery et al., 1984; Walden et al., 1981). A study with adults with normal hearing (Massaro et al., 1993) showed that perception of trained nonsense syllables and isolated words improved, but there was not adequate evidence for generalization.

The presupposition that analytic training would be more efficient and effective has been shown in contemporary perceptual learning research to have failed to take into account how trainees actually carry out explicit phoneme identification. Research suggests that phoneme identification training often leads to overlearning stimulus-specific features and/or adopting strategies specific to the task. This mode of learning has been referred to as reflective learning (Chandrasekaran et al., 2014). Reflective learning may involve trying to generate and apply explicit rules, which in the case of speech, is typically not an effective strategy for achieving generalization. Importantly, explicit phoneme identification is a metalinguistic task that is not part of ordinary speech recognition (Hickok & Poeppel, 2000). So the rationale of analytic training—learning to identify phonemes in order to generalize to recognizing words—seems to have been incorrect.

Twentieth century synthetic training research recognized the need for ecologically valid stimuli. It also presupposed that the trainee needed to learn higher-level
strategies such as the use of semantic context. However, training with sentence stimuli posed technical challenges for how to present the stimuli, and how to give feedback. One solution was a procedure in which an experimenter reads text for the trainee to lipread, repeating each phrase until the trainee repeats it back *verbatim* (DeFilippo, 1984; Matthies & Carney, 1988), a technique referred to as *Connected Discourse Tracking*. The number of words correct per minute was the trainee’s score. The efficacy of the approach was difficult to evaluate. It does not provide adequate experimental control to attribute increased correct word rate to learning to lipread more accurately rather than, for example, to learning the vocabulary used in the text or learning subtle cues from the live talker (Tye-Murray & Tyler, 1988).

Our early training studies used sentence presentation under computer control in the context of investigating vibrotactile aids to lipreading (e.g., L. E. Bernstein et al., 1991; Eberhardt et al., 1990). Participants with normal hearing lipread sentences with or without accompanying vibrotactile speech and received feedback in the form of the correct response printed on a computer screen. Lipreading training with this type of printed feedback but without vibrotactile speech was not very effective.

Using the computer-controlled stimuli and correct-response feedback in a study with younger adults that sought to determine whether practice would reduce differences between good lipreaders who were congenitally deaf versus ones with normal hearing (L. E. Bernstein et al., 2001), the participant groups remained reliably different (*p* < 0.05), with deaf participants maintaining their advantage. There was only a small although reliable (on the order of 2 percentage points, *p* < 0.05) increase in VO sentence scores across training sessions and groups.

We were not alone in early attempts to present recorded connected speech stimuli under computer control during training (e.g., Boothroyd, 2010). However, to our knowledge, no research was reported that applied any feedback techniques during training other than displaying the correct response. Below, we discuss why the design of external feedback may be a critical element for successful lipreading training.

**Summary of the 20th Century Training Research**

Analytic and synthetic training approaches delivered modest, if any, gains in lipreading. Attempts to obtain efficient and effective training through an analytic task never resulted in substantial generalization to untrained materials. Synthetic training was limited by technology and inadequate knowledge of perceptual learning and its requirements.

**Contributions From Visual Perceptual Learning Research for Solving the Problem of Achieving Effective Lipreading Training**

Following our earlier lipreading training studies, we carried out research on the neural bass for lipreading, positing that VO speech—speech as a visual stimulus—must be processed extensively through the visual system (e.g., L. E. Bernstein et al., 2008, 2011; L. E. Bernstein & Liebenthal, 2014). The work suggested, and subsequent research confirmed (e.g., Nidiffer et al., 2021; Venezia et al., 2017), that VO speech is processed *qua* speech in areas of the visual cortical hierarchy. In turn, that work motivated us to consider of perceptual learning from the perspective of vision science, which has led the way in perceptual learning research.

Perceptual learning is defined on the behavioral level as “long-lasting changes to an organism’s perceptual system that improve its ability to respond to its environment” (Goldstone, 1998). Accepted theoretical accounts of perceptual learning that are expected to apply across the different sensory-perceptual systems posit bottom-up sensory representations, top-down internal feedback, and interactions with external feedback (e.g., Ahissar et al., 2009; F. G. Ashby & Maddox, 2011; Ashby & Valentin, 2017; Dosher & Lu, 2017; Friston, 2005; Hochstein & Ahissar, 2002; Watanabe & Sasaki, 2015). Internal feedback is feedback from higher level to lower level areas within the perceiver’s hierarchically organized cortical networks. Internal feedback can be sufficient for perceptual learning when stimuli are obvious (J. Liu et al., 2012), but external feedback is likely required in order for adults to learn difficult stimuli (Ashby & Maddox, 2011; Chandrasekaran et al., 2014; J. Liu et al., 2010; Z. Liu et al., 2010). Visual speech is difficult to learn as an adult, and external feedback is generally viewed as necessary.

**Reverse Hierarchy Theory**

Earlier, we outlined evidence that within-viseme perceptual distinctions can be made by adults with normal hearing (Files et al., 2013). This evidence is critical in relationship to perceptual learning of visual speech and to the role of feedback in perceptual learning: Perceptual learning involves learning information that is available in bottom-up sensory representations (Ahissar et al., 2009; Nahum et al., 2010). The use of feedback to guide learning information that is transduced by the senses and represented at the level of the cortex but is not used to carry out a specific perceptual task is explained within the reverse hierarchy theory (RHT) of perceptual learning (Ahissar et al., 2009; Nahum et al., 2010). The hierarchy in the RHT is the hierarchical organization of sensory-perceptual systems of the cerebral cortex that represent stimuli through bottom-up representations of increasing...
generality at higher cortical levels. The RHT posits that, whenever possible, perceptual tasks rely on the highest possible available neural representations, that is, the highest relevant generalizations or categories. When a perceptual task cannot be carried out based on higher-level cortical representations, external feedback may be needed to guide attention to stimulus information that is represented at lower levels of the cortical hierarchy. In the case of lipreading, perceptual learning of sublexical visual speech information requires accessing stimulus representations that offer discriminable speech information. These representations of discriminable information can then be mapped to new higher-level speech features or categories. That is, RHT posits that the information needed for perceptual learning is represented within the cortical hierarchy even though the information may not be used ordinarily for distinguishing categories for a particular task. The task for lipreading is word recognition.

To be clear, our review of lipreading above explained that there is evidence that adults have access to visual phonetic detail to discriminate phonemes within putative visemes, and that they can access that information to select a target word from between two isolated words with the same visemes. Furthermore, they may be limited in using the information to isolate words in the mental lexicon during an open set task such as sentence lipreading. RHT suggests that training should be carried out with external feedback that supports access to the available sensory-perceptual representations for learning new categories for carrying out the task of lipreading words.

Visual perceptual learning research also suggests that in order for external feedback to be effective, it needs to be contingent on the perceptual errors that are made during training, rather than merely providing knowledge about the correct response, regardless of the response (Ashby & Vucovich, 2016). Even if feedback is given for only correct words, the feedback lacks contingency with regard to words that were partially correct versus entirely incorrect. Correct-only feedback provides no guidance for increasing the use of sublexical speech information that the participant can discriminate but has not learned to use in recognizing words. In fact, feedback for partially correct responses may be one key to inducing perceptual learning of visual speech. This suggestion strongly contrasts with the synthetic methods of lipreading training, which present whole sentence or phrase stimuli, collect a response, and then inform the participant what the talker just said or confirm the response words that were correct.

Application of the RHT of Perceptual Learning to Lipreading Training

We recently applied the theoretical perspective from the RHT in developing an approach to lipreading training that uses sentence stimuli for their ecological validity and feedback contingency based on phoneme-level scoring to support sublexical visual speech perceptual learning (L. E. Bernstein et al., 2021). Thus, the training incorporated the phoneme and connected speech linguistic levels that were the focus of analytic and synthetic training paradigms, respectively (Blamey & Alcantara, 1994; DeFilippo, 1988; Lesner et al., 1987; Montgomery et al., 1984; Walden et al., 1981). Furthermore, the design of the training, which used phoneme-level scoring and feedback contingent on the response, was based on the contemporary view of perceptual learning outlined in the previous section.

We developed a method for automatic phoneme-level open set scoring that quantifies phonemic misperceptions on a continuous scale of perceptual dissimilarity (L. E. Bernstein et al., 1994). The scoring procedures were used to generate feedback during training. We compared learning across three different groups of younger adults with normal hearing (N = 57) who trained in six sessions. One training group (Sentence Group) received traditional synthetic feedback in the form of the entire printed stimulus sentence following two attempts to type what the talker had said. Another group received word-level feedback: After their first attempt to type what the talker said, they saw printed on their screen their correct response words as well as the correct words for incorrect response words that were perceptually similar to the correct response (Word Group). That is, the feedback contingency for incorrect words was based on the response words’ perceptual similarity to correct words. The third group (Consonant Group) received consonant-level feedback in addition to word feedback: After their first attempt to type what the talker said, they saw printed on their screen their correct response words and the consonants in the stimulus words that were incorrectly perceived but resulted in responses that were perceptually similar to correct word responses. That is, the Consonant Group’s feedback was contingent on consonants that were perceptually similar to the stimuli. Thus, the Sentence Group’s feedback was not contingent on their response, whereas the Word and Consonant Groups’ feedback was, but there were differences in the context of the feedback between the Consonant and Word Groups. Word feedback implicitly offered information about perceptual errors, but consonant feedback did so more explicitly.

Indeed, the Consonant Group learned reliably more than did the other two training groups and more than an untrained Control Group. Importantly, the Consonant Group’s learning generalized to untrained sentences in VO, AV, and AO stimulus conditions with the audio stimuli in speech-shaped noise (p < .05). Between pre- and posttraining tests, the Consonant Group’s mean percent words correct scores increased by 9.2 percentage points.
Application of an Auditory Perceptual Learning Paradigm to Lipreading Training

Several previous studies on learning to recognize vocoded speech (Davis et al., 2005; Hervais-Adelman et al., 2008; Sohoglu & Davis, 2016; Sohoglu et al., 2012) have been interpreted through predictive coding theory (Friston, 2005, 2010). One of the features of predictive coding theory is the role it specifies for top-down internal predictions of stimulus information and the updating of those predictions, when they do not match bottom-up stimulus representations. Predictive coding theory invokes bottom-up error signals to explain perceptual learning, which is considered to be the learning of more accurate predictions.

Research carried out by Davis and colleagues demonstrated that when a stimulus is printed first, followed by a vocoded AO training stimulus, learning by adults with normal hearing generalizes to untrained words or sentences more than if the printed text of the stimulus follows (Davis & Johnsrude, 2007; Davis et al., 2005). Prior print also increases word clarity judgments (Sohoglu & Davis, 2016; Sohoglu et al., 2012).

In a neurophysiological study (Sohoglu & Davis, 2016), in which prior lexical knowledge and stimulus clarity (number of noise bands) of vocoded speech were manipulated, cortical source reconstructions showed that both prior lexical knowledge and stimulus clarity modulated activity in the temporal gyrus posterior to Heschl’s gyrus in the auditory cortex. The authors interpreted their results as evidence that printed text prior to vocoded speech generates lexical candidate predictions (top-down feedback) for the upcoming vocoded speech stimulus, and that the prediction is confirmed or disconfirmed by the bottom-up sensory representations of the stimulus, with subsequent updating of top-down predictions of sublexical speech categories. That is, their results were consistent with predictive coding theory.

An important point to emphasize about the vocoder research outlined above is that the prior printed text initiates internal top-down feedback to representations in an expert system that can perceive all of the phonemes that are needed to recognize words. Another observation is that the top-down feedback of printed words targets auditory cortical areas. In light of the visual speech processing pathways found in neural research on lipreading (L. E. Bernstein et al., 2017; L. E. Bernstein & Liebenthal, 2014; Nidiffer et al., 2021; Venezia et al., 2016, 2017), we hypothesized that initiating top-down internal feedback that automatically targets auditory speech representations will interfere with learning to lipread more accurately, or at the least, will not be helpful (L. E. Bernstein et al., submitted). The neuroimaging research on visual speech shows that visual phonetic processing engages high-level visual areas, possibly to the level of word forms, so auditory cortex targets for phonetic representations would not seem to be useful targets for lipreading training.

We carried out a behavioral experiment to investigate the effects of prior printed text during lipreading training (L. E. Bernstein et al., submitted). Prior orthographic text was predicted to interfere with VO perceptual learning in a task that had been used previously in studying vocoder and visual speech perceptual learning (L. E. Bernstein et al., 2013, 2014; Eberhardt et al., 2014). The training task was to learn, on each of four training days, to uniquely identify a different set of 12 novel nonsense objects (images) that were paired with 12 novel nonsense words. We assigned 88 younger adults with normal hearing to different training groups or to an untrained control group.

During novel word training, one of the groups (Word Group) saw each novel word printed before the VO speech stimulus. Another group (Consonant Group) saw the printed word with its vowels removed prior to the VO speech stimulus. We reasoned that by withholding vowel information from the printed text, participants would be less likely to initiate top-down lexical feedback to their highly expert auditory representations and would instead shift toward more sustained or focused attention on the visual speech during training. We also reasoned that because the task was to learn words, and the feedback following novel image selection during training trials was based on image selection not consonants, participants would not benefit from using a reflective learning strategy (Chandrasekaran et al., 2014). A VO Control Group received no prior printed information during their training. Finally, a Vocoder Group received vocoded acoustic speech followed by VO speech. The Vocoder Group was a

---

3A vocoded speech signal has been input to one or more filters that output the filter signal level/s, which is/are then used to modulate a sine wave or noise band. The output signals are added prior to presentation to the listener.
control for the reduced information that the Consonant Group received during training.

Uniquely, the Consonant Group increased its scores \((p < .05)\) between pre- and posttraining tests of untrained VO sentences spoken by a talker who was not seen during training. After only four training sessions of about 20 min each on different days, the Consonant Group improved \((p < .05)\) its open set identification of words in VO sentences by an average of approximately 4 percentage points. The pattern of performance during training by the Word Group suggested that they relied on the prior printed words to learn the word-picture pairs and failed to learn to lipread the VO spoken training words. They appeared to ignore the visual speech in favor of the printed words. This study unfortunately did not test using pre- and posttraining AV and AO sentences. However, an ongoing clinical trial with this paradigm for older adults who have acquired hearing loss is carrying out extensive pre- and posttraining tests with VO, AO, and AV stimuli.

**Summary and Conclusions About Lipreading Training**

There have been numerous attempts to achieve effective lipreading training paradigms. Those of the 20th century focused on how knowledge about the structure of language (e.g., phonemes, words, and sentences) might be deployed to obtain efficient and effective training. Relative to the present, there was limited knowledge about the roles of internal and external feedback in perceptual learning. There was also much less knowledge about how the training task may affect perceptual learning.

We approached our recent designs for lipreading training by considering current theories about perceptual learning and the functional neuroanatomy of lipreading. In training studies with very different paradigms that were applied with younger adults who had normal hearing, we obtained evidence for reliable VO training effects, as well as evidence of generalization to untrained AV speech recognition in noise. The results of both training studies suggest that the conditions for effective lipreading training are very specific. External feedback is needed that directs internal feedback to sublexical stimulus information that likely can already be discriminated. The externally provided feedback can precede or follow the training stimulus, depending on the paradigm. However, full lexical information prior to the stimulus can impede lipreading learning. More research is needed to expand the available paradigms for training and to extend our understanding of the specific conditions that engage perceptual learning with generalization. Currently, we are carrying out lipreading training studies with older adults who have acquired hearing loss.

**Comments About Auditory Speech Training**

A reviewer suggested that we briefly address AO speech perception training for adults with acquired hearing loss. In comparison with research on lipreading training, there is a much larger contemporary literature on this topic. Thus, a brief commentary necessarily fails to do justice to the full effort in this area. There are some available reviews for the interested reader (e.g., Henshaw & Ferguson, 2013; Lawrence et al., 2018). The reviews describe the efficacy of auditory speech training as modest at best, with little evidence for generalization to untrained tasks and/or stimuli. Difficulties in achieving generalization have been countered by the proposal that interventions should instead target cognitive abilities such as working memory and attention (Ferguson et al., 2014; Rönberg et al., 2013, 2019), inasmuch as such mechanisms may be deployed to compensate for perceptual difficulties. There is also a very large literature on the relationship between cognitive abilities and auditory or speech recognition abilities, with several reviews available (e.g., Akeroyd, 2008; Dryden et al., 2017).

In their recent review and meta-analysis of auditory and cognitive training, Lawrence et al. (2018) concluded that “Overall certainty in the estimation of effect was “low” for auditory training and “very low” for cognitive training.” They advocated for “high-quality RCTs (randomized controlled trials)” to better judge efficacy. Recently, Henshaw and colleagues (Henshaw et al., 2021) reported on a working memory training program that was administered in an RCT carried out with adults who were hearing aid users. Their study obtained no evidence that the training generalized to speech recognition in noise.

An examination of the nine studies that met inclusion criteria for the Lawrence et al. (2018) review and meta-analysis (see their table 1) lists five “auditory [speech] training” studies. Here, we note that the methods applied in those studies followed from the traditional analytic or synthetic training approaches. For example, the computer-based LACE program (Sweetow & Sabes, 2006) is described by its authors as “analytic,” although the participants listened to degraded connected speech. Following each response, participants received feedback informing them of the completely correct response. We discussed above that synthetic training on connected speech with low contingency feedback—the same feedback information regardless of the response—is unlikely to be highly effective. In another of the reviewed studies, one carried out by Ferguson et al. (2014), participants trained on a phoneme discrimination task that used synthesized syllables, again with “low” effects. We discussed above how analytic training may offer trainees the opportunity to learn explicit phoneme identification but without generalization.
beyond the task, as was reported by the Ferguson et al. study.

A recent line of research that used exclusively auditory speech analytic training with frequently occurring words also resulted in no generalization to untrained materials. The research, which culminated in an RCT (Burk & Humes, 2008; Humes et al., 2009, 2018), focused on word-level training with frequently occurring words presented in a forced choice paradigm. Participants were adults with mild-to-moderate hearing loss. The failure of this approach might be attributable to using the highly familiar words, encouraging superficial attention to the stimuli, or the use of a forced choice task, encouraging task learning or guessing. The correct word feedback may also have reduced learning.

It might be thought that we would argue for translating our paradigms to the auditory domain. That is not our position. A range of modality- and experience-specific factors need to be taken into account in developing training paradigms. A lifetime of auditory and visual speech processing is expected to result in auditory and visual neural modality-specific and behavioral expertise-specific effects on the individual. As a consequence, what is effective for training lipreading in a particular population or individual may be ineffective for training auditory or AV speech recognition in that population or individual.

For example, when we applied our novel word learning paradigm with adults who had normal hearing in comparison with adults with prelingual deafness and late acquired cochlear implants (L. E. Bernstein et al., 2014), we obtained a different pattern of results across the two groups. During training in a crossover design, participants in each group were trained first with AO speech (i.e., vocoded for the adults with normal-hearing, cochlear implant for the prelingually deaf adults) and then with AV speech, or vice versa. Novel word tests, which immediately followed each training session, were always carried out AO. The results showed that when the training was AV the adults with CIs had difficulty learning to recognize the novel auditory word stimuli. Their lifelong reliance on visual speech appeared to impede their auditory word learning. The adults with normal hearing did not demonstrate a difference in word learning as a function of the AO versus AV training stimulus modality.

Also, the study resulted in evidence from pre- and posttraining tests of auditory phoneme identification with CVCV-CVC stimuli that the deaf adults improved their initial AO consonant perception, whereas those with normal hearing improved their medial AO consonant perception of the vocoded speech. The deaf participants’ learning was consistent with the higher intelligibility of initial visual consonants than medial or final consonants, whereas the learning by participants with normal hearing was consistent with the diverse information in vowel transitions into and out of medial consonants in acoustic speech, rendering the medial consonants more intelligible. That is, the participants’ lifelong perceptual experience biased what they learned.

A consideration for auditory training, as we suggested above for lipreading training, is that support may be needed for accessing the representations of phonetic cues that can be heard but have not been learned—are not used in perceiving speech. Specifically, individuals vary in their use of acoustic-phonetic cues (Broersma & Scharenborg, 2010; Haggard et al., 1970; e.g., Hazan & Rosen, 1991; Kapnoula et al., 2017; Nittrouer et al., 2015; Roque et al., 2019; Van Tasell et al., 1987; Zlatin, 1974). Cues may be audible, but the individual listener with hearing loss may not have learned to use that information (Broersma & Scharenborg, 2010; Cutler et al., 2004; Lecumberri et al., 2010). The sublexical cues that they have learned to rely on may be less useful than other cues that are available for recognizing words in degraded speech (Lowenstein et al., 2012; Nittrouer et al., 2015). Another example might be using the talker’s vocal characteristics to separate the target from the background. Training might be useful, if the listener does not attend to that information (Cooke et al., 2008; Hoen et al., 2007; Mattys et al., 2012).

We think that there are different avenues to potentially achieving effective auditory speech training for adults with acquired hearing loss. The design of training may need to be very specific, as seems to be the case for lipreading. Training must take into account the trainee’s current speech processing mechanisms, the availability of perceptible but unlearned speech information, and training task design. Another consideration for auditory training may be related to the stability of auditory speech perception in the context of stimulus noninvariance (see below). Stability in lipreading may also be related to stimulus selection bias (see below).

Implications for AV Training and Hearing Aid Signal Processing Research

It is clear that the perceptual integration of auditory and visual speech stimuli is more than the sum of unisensory auditory and visual speech perception (e.g., Grant & Seitz, 1998; Opoku-Baah et al., 2021; Ross et al., 2007). The mechanisms of AV speech integration and how and why they differ from unisensory AO and VO processing is far beyond the scope of this review article. However, there are several issues regarding AV speech that we think are important to mention in the context of lipreading training for improving AV speech recognition in adults with acquired hearing loss. Specifically, we comment below on AV training in relationship to stimulus selection bias and in relationship to the stability of speech perception.
AV Training and Stimulus Selection Bias

One of the challenges of training with AV speech is that participants have had lifelong experience with this type of stimulus. Prior experiences processing certain of the information in complex stimuli can bias attention away from other stimulus information, while also reducing the ability to learn the other stimulus information. This is an example of “stimulus selection bias,” whereby stimulus information can be perceived but is not processed (Awh et al., 2012).

A study by Montgomery et al. (1984) implicitly addressed stimulus selection bias. The researchers conducted a 10-session training program in which the task for 24 adults with aided mild-to-moderate hearing loss and mean age 39 years was to recognize words in connected speech. The live AV speech stimuli were processed, so that weaker sounds were removed in order to encourage the participants to attend to the visual information. Much of the consonantal information was removed. A control group received an AO rehabilitation program. The expectation was that removal of the consonantal information would shift attention to the visual speech. There was some evidence that AV training was quantitatively more effective for AV sentences across pre- and posttraining tests. However, statistical analysis that took into account pre-training scores failed to show a differential effect of AO versus AV training. The difference between pre- and post-training words correct in AV sentences was about 4 percentage points. Thus, although there was, potentially, information to be learned from the visual speech, the participants failed to do so at a level that was beyond that of AO training.

Stimulus selection bias may also have been responsible for the findings in a recent study that examined adults’ eye gaze while viewing talking faces (Rennig et al., 2020). Under noisy AV conditions, all of the younger adult participants, who had normal hearing, shifted to consistently looking at the talker’s lower face. Furthermore, their speech perception accuracy was not predicted by their eye gaze during the noisy AV speech. Their speech perception was predicted by their fixation during clear speech. That is, those participants who were biased to look more at the lower face in clear conditions were more likely to have learned more visual speech information, which supported their AV speech recognition when the acoustic speech was noisy.

An important explanation for stimulus selection bias away from visual speech may be that faces disclose socially significant competing information such as the talker’s identity, gaze direction, emotion, and affect. That is, as a visual stimulus, speech overlaps with social face stimuli, which may be more salient to the perceivers than are visual speech. Stimulus information that is more salient and that overlaps spatially and temporally with less salient information impedes perceptual learning of the less salient information (Hammer et al., 2015).

AV Training and the Stability of Speech Perception

There is also research on AV training that supports the conclusion that speech processing mechanisms are biased toward long-term stability. Results from so-called “phonetic recalibration” experiments (Vroomen & Baart, 2012) support the conclusion that AV speech stimuli are more likely to induce short-term effects than long-term perceptual learning.

In recalibration experiments with AV speech, listeners label an ambiguous sound (/?/) half-way between two phonemes such as /b/ and /d/. If they previously experienced the ambiguous token with a visual /b/, then the AO stimulus is identified as “b” more than “d” and vice versa (Bertelson et al., 2003). However, recalibration lasts only about six trials (Vroomen & Baart, 2012). The effect may also be relatively speaker- or token-specific (van der Zande et al., 2014), and it has been shown to be earspecific (Keetels et al., 2015). Notably, Keetels et al. (p. 125) remark that, “Our results highlight that speech recognition rather prefers to maximize stability over its lifelong experience, and that learning is thus very specific to the situation encountered.” If this is a general property of perception of AV speech, then the usefulness of training with AV speech may be extremely limited for achieving lasting effects.

On the other hand, individuals vary widely with regard to the extent to which they integrate auditory and visual information (Mallick et al., 2015; McGurk & MacDonald, 1976; Odegaard & Shams, 2016). There may be a role for AV training that targets the ability to integrate auditory and visual stimuli. However, very little research has focused on training to increase integration. A suggestion from a study with nonspeech stimuli is that temporal alignment is important for achieving binding of auditory and visual stimuli, and manipulation of temporal alignment can be used to modify integration, at least over relatively short periods of time (Odegaard et al., 2017). Furthermore, design of training with AV speech to achieve long-lasting perceptual learning for integration is an essentially unexplored area for adults with acquired hearing loss.

In summary, this section suggests that AV training paradigms would need to overcome biases of highly practiced conservative speech recognition systems. Long-term perceptual learning likely requires very specific training conditions in order to overcome system preferences for stability.
Lipreading in Relationship to Acoustical Signal Processing

Last, we think there may be benefits attached to considering how visual speech might be important for testing hearing aid algorithms. There is a large literature on hearing aid algorithms that we cannot possibly do justice to here. This section only aims to point out that in actual listening situations, when the talker can be seen, signal processing algorithms may have different effects on perception than in the laboratory.

An acknowledged difficulty in developing and testing acoustic signal processing algorithms for hearing aids is carrying out ecologically valid testing (Kollmeier & Kiessling, 2018). Although noise is a common listening condition that is frequently used in testing, visual speech is not commonly used. Visible speech is however too difficult to introduce into the laboratory.

In addition to increasing ecological validity by using stimuli that are more like speech communication in social settings, research suggests that AV speech may shift the importance of acoustic frequency bands. Bernstein and colleagues (J. G. W. Bernstein et al., 2020) carried out a study to determine the importance of different spectral regions for adults with normal hearing (N = 4) or hearing loss (N = 8) when speech was AO versus AV. When speech was AV, low-frequency information increased in importance relative to higher frequency information. When speech was AO, the importance function was relatively flat across frequency bands. These results imply that shifts in frequency importance due to visual speech should be taken into account in developing algorithms for devices that are designed to improve the ability of users to comfortably socialize.

AV effects may be particularly important to take into account in relationship to frequency lowering algorithms that are intended to improve speech understanding. Frequency lowering or compression schemes aim to redistribute spectral information to avoid cochlear dead zones. “Frequency lowering schemes have been advocated in the literature and have been applied in commercial hearing devices without a clear prove [sic] of a benefit for listeners” (Kollmeier & Kiessling, 2018). Hearing aid companies started offering frequency lowering in 2006 (Alexander, 2016). Early algorithms caused distortion by overlaying higher frequency spectral bands on lower frequency bands and were relatively ineffective at improving perception of AO sentences in quiet or in noise (Alexander, 2013; Bruno et al., 2021; Miller et al., 2016; Yakumina & Nam, 2021). Some frequency compression algorithms appear to be more effective in quiet (Alexander & Rallapalli, 2017) and others in noise for some users but not everyone (Bohnert et al., 2010; Hopkins et al., 2014; Shehorn et al., 2018). Some preliminary evidence was reported showing that frequency lowering may benefit users with severe or profound hearing loss when listening with AV speech but not AO speech (Sakamoto et al., 2000).

In their review of hearing aid effectiveness, Kollmeier and Kiessling (2018) suggest with regard to frequency lowering that, “only a limited effect is expected in the majority of hearing-impaired listeners and the vast variation of natural acoustical environments. Moreover, an extensive training program would be required which can hardly be supplied in laboratory studies. This may contribute to the fact that no clear advantage of this kind of processing has been demonstrated in the literature so far” (p. S17). However, frequency lowering may be more effective and trainable in the context of visual speech information. Specifically, the evidence for recalibration of ambiguous acoustic speech by visual speech (Vroomen & Baart, 2012) may be useful in devising schemes for moving frequency information into audible portions of the spectrum for individual listeners. Specifically, listeners may be able to deploy recalibration while they are in face-to-face listening conditions, even if the effects fade quickly when the talker is no longer visible. Even short-term adaptation to acoustic speech signals in social contexts in which the talker can be seen would seem to be desirable, if intelligibility can be significantly enhanced. However, users of a hearing aid algorithm devised specifically for AV listening conditions would need to understand when an algorithm would provide good or poor results and switch algorithms when appropriate.

Summary and Conclusions

Visual speech information is ubiquitous in face-to-face social settings. Lipreading offers opportunities for improving AV speech recognition by individuals with acquired hearing loss, particularly when speech is in noisy environments. Research on lipreading was an integral part of the field of audiology in the 20th century. However, in the 21st century, since the improvement of hearing aids, lipreading has not been given the attention it deserves in research and in the clinic. Maximizing speech recognition with a hearing loss still requires use of visual speech information.

This review article sought to reinvigorate interest in lipreading training as a treatment option. Because adults with acquired hearing loss are likely to be poor lipreaders, training may be useful for them to improve their lipreading and their AV speech recognition in noise. Theoretical bases for developing new training paradigms were discussed, and recent examples of successful training were described. An open area for training research concerns discovering the conditions that promote perceptual learning as opposed to perceptual stability. Another area for
research is how visual speech may interact with signal processing algorithms in hearing aids.

Acknowledgments

Work on this study was supported by the National Institutes of Health/National Institute on Deafness and Other Communication Disorders (R56 DC016107 [Bernstein, PI], R21 DC014523 [Bernstein, PI], and R44 DC015418 [Eberhardt, PI]).

References

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. Philosophical Transactions of the Royal Society B, 364(1515), 285–299. https://doi.org/10.1098/rstb.2008.0253

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. International Journal of Audiology, 47(Suppl. 2), S53–S71. https://doi.org/10.1080/14992020802301142

Alexander, J. M. (2013). Individual variability in recognition of frequency-lowered speech. Seminars in Hearing, 34(3), 253–254. https://doi.org/10.1055/s-0033-1348021

Alexander, J. M. (2016). 20Q: Frequency lowering ten years later - New technology innovations. AudiologyOnline, Article 18040.

Alexander, J. M., & Rallapalli, V. (2017). Acoustic and perceptual effects of amplitude and frequency compression on high-frequency speech. The Journal of the Acoustical Society of America, 142(2), 908–923. https://doi.org/10.1121/1.4997938

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. Annals of the New York Academy of Sciences, 1224(1), 147–161. https://doi.org/10.1111/j.1749-6632.2010.05874.x

Ashby, F. G., & Valentin, V. V. (2017). Chapter 7 - Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), Handbook of Categorization in Cognitive Science (2nd ed., pp. 157–188). Elsevier. https://doi.org/10.1016/B978-0-40-101107-2.00007-5

Ashby, F. G., & Vucovich, L. E. (2016). The role of feedback contingency in perceptual category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 42(11), 1731–1746. https://doi.org/10.1037/xlm0000277

Auer, E. T., Jr. (2002). The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness. Psychonomic Bulletin & Review, 9(2), 341–347. https://doi.org/10.3758/BF03196291

Auer, E. T., Jr., & Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. The Journal of the Acoustical Society of America, 102(6), 3704–3710. https://doi.org/10.1121/1.420402

Auer, E. T., Jr., & Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. Journal of Speech, Language, and Hearing Research, 50(5), 1157–1165. https://doi.org/10.1044/1092-4388(2007/080)

Awh, E., Belopolsky, A., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. Trends in Cognitive Sciences, 16(8), 437–443. https://doi.org/10.1016/j.tics.2012.06.010

Bernstein, J. G. W., Venezia, J. H., & Grant, K. W. (2020). Auditory and auditory-visual frequency-band importance functions for consonant recognition. The Journal of the Acoustical Society of America, 147(5), 3712–3727. https://doi.org/10.1121/10.0001301

Bernstein, L. E. (2012). Visual speech perception. In E. Vatikiotis-Bateson, G. Bailly, & P. Perrier (Eds.), Audiospatial speech processing (pp. 21–39). Cambridge University. https://doi.org/10.1017/CBO9780511843891.004

Bernstein, L. E. (2018). Response errors in females’ and males’ sentence lipreading necessitate structurally different models for predicting lipreading accuracy. Language Learning, 68(S1), 127–158. https://doi.org/10.1111/langu.12281

Bernstein, L. E., Auer, E. T., Jr., & Eberhardt, S. P. (2021). During lipreading training with sentence stimuli, feedback controls learning and generalization to audiovisual speech in noise. American Journal of Audiology. Advance online publication. https://doi.org/10.1044/2021_AJA-21-00034

Bernstein, L. E., Auer, E. T., Jr., Eberhardt, S. P., & Jiang, J. (2013). Auditory perceptual learning for speech perception can be enhanced by audiovisual training. Frontiers in Neuroscience, 7, 34. https://doi.org/10.3389/fnins.2013.00034

Bernstein, L. E., Auer, E. T., Jr., & Tucker, P. E. (2001). Enhanced speechreading in deaf adults: Can short-term training/practice close the gap for hearing adults? Journal of Speech, Language, and Hearing Research, 44(1), 5–18. https://doi.org/10.1044/1092-4388(2001/001)

Bernstein, L. E., Auer, E. T., Jr., Wagner, M., & Ponton, C. W. (2008). Spatiotemporal dynamics of audiovisual speech processing. NeuroImage, 39(1), 423–435. https://doi.org/10.1016/j.neuroimage.2007.08.035

Bernstein, L. E., Demorest, M. E., Couter, D. C., & O’Connell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. The Journal of the Acoustical Society of America, 90(6), 2971–2984. https://doi.org/10.1121/1.401771

Bernstein, L. E., Demorest, M. E., & Eberhardt, S. P. (1994). A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus-response alignment. The Journal of the Acoustical Society of America, 95(6), 3617–3622. https://doi.org/10.1121/1.4069930

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. Perception & Psychophysics, 62(2), 233–252. https://doi.org/10.3758/BF03205546

Bernstein, L. E., Eberhardt, S. P., & Auer, E. T. (2002). Novel word learning of visual speech versus vocoded speech is affected differently by word versus phoneme feedback type. Manuscript submitted for publication.

Bernstein, L. E., Eberhardt, S. P., & Auer, E. T., Jr. (2014). Audiovisual spoken word training can promote or impede auditory-only perceptual learning: Prelingually deafened adults with late-acquired cochlear implants versus normal hearing adults. Frontiers in Psychology, 5, 934. https://doi.org/10.3389/fpsyg.2014.00934

Bernstein, L. E., Eberhardt, S. P., & Demorest, M. E. (1989). Single-channel vibrotactile supplements to visual perception of intonation and stress. The Journal of the Acoustical Society of America, 85(1), 397–405. https://doi.org/10.1121/1.397690

Bernstein, L. E., Eberhardt, S. P., Jiang, X., Riesenhuber, M., & Auer, E. T. (2017). The representation of lipread words in posterior temporal cortex studied using an fMRI-rapid adaptation paradigm and functional localizers. Paper presented at the Neuroscience 2017, Washington, DC.

Bernstein, L. E., Jiang, J., Pantazis, D., Lu, Z. L., & Joshi, A. (2011). Visual phonetic processing localized using speech and nonspeech
face gestures in video and point-light displays. Human Brain Mapping, 32(10), 1660-1676. https://doi.org/10.1002/hbm.21139

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. Frontiers in Neuroscience, 8, 386. https://doi.org/10.3389/fnins.2014.00386

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk after-effect. Psychological Science, 14(6), 592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Blamey, P. J., & Alcantara, J. I. (1994). Research in auditory training. In J. Gagne & N. Tye-Murray (Eds.), Research in audiological rehabilitation: Current trends and future directions (pp. 161–191). Academy of Rehabilitative Audiology.

Bohnert, A., Nyffeler, M., & Keilmann, A. (2010). Advantages of a non-linear frequency compression algorithm in noise. European Archives of Oto-Rhino-Laryngology, 267(7), 1045-1053. https://doi.org/10.1007/s00405-009-1170-x

Boothroyd, A. (2010). Adapting to changed hearing: The potential role of formal training. Journal of the American Academy of Audiology, 21(9), 601–611. https://doi.org/10.3766/jaaa.21.9.6

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 43(3), 647–677. https://doi.org/10.1080/14640749108400991

Broersma, M., & Scharenborg, O. (2010). Native and non-native listeners’ perception of English consonants in different types of noise. Speech Communication, 52(11–12), 980–995. https://doi.org/10.1016/j.specom.2010.08.010

Bruno, R., Freni, F., Portelli, D., Alberti, G., Gazia, F., Meduri, A., Galletti, F., & Galletti, B. (2021). Frequency-lowering processing to improve speech-in-noise intelligibility in patients with age-related hearing loss. European Archives of Oto-Rhino-Laryngology, 279, 3697–3706. https://doi.org/10.1007/s00405-020-06431-8

Burk, M. H., & Humes, L. E. (2008). Effects of long-term training on aided speech-recognition performance in noise in older adults. Journal of Speech, Language, and Hearing Research, 51(3), 759–771. https://doi.org/10.1044/1092-4388(2008/054)

Chandrasekaran, B., Yi, H. G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. Psychonomic Bulletin & Review, 21(2), 488–495. https://doi.org/10.3758/s13423-013-0501-5

Clouser, R. A. (1977). Relative phoneme visibility and lipreading performance. Volta Review, 79(1), 27–34.

Conrad, R. (1977). Lip-reading by deaf and hearing children. The British Journal of Educational Psychology, 47(1), 60–65. https://doi.org/10.1111/j.2044-8279.1977.tb03001.x

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. The Journal of the Acoustical Society of America, 123(1), 414–427. https://doi.org/10.1121/1.2804952

Cutler, A., & Jesse, A. (2021). Word stress in speech perception. In J. S. Pardo, L. C. Nyagaard, R. E. Remez, & D. B. Pisoni (Eds.), The handbook of speech perception (pp. 239–265). https://doi.org/10.1007/9781119118406.ech9

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. The Journal of the Acoustical Society of America, 116(6), 3668–3678. https://doi.org/10.1121/1.1810292

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. Hearing Research, 229(1–2), 132–147. https://doi.org/10.1016/j.heares.2007.01.014

Davis, M. H., Johnsrude, I. S., Hervaïs-Adelman, A., Taylor, K., & McGgettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. Journal of Experimental Psychology: General, 134(2), 222–241. https://doi.org/10.1037/0096-3455.134.2.222

DeFilippio, C. L. (1984). Laboratory projects in tactile aids to lip-reading. Ear and Hearing, 5(4), 211–227. https://doi.org/10.1097/00003446-198407000-00006

DeFilippio, C. L. (1988). Tracking for speechreading training. Volta Review, 90(5), 215–239.

Dias, J. W., McClaskey, C. M., & Harris, K. C. (2021). Audiospatial speech is more than the sum of its parts: Auditory-visual superadditivity compensates for age-related declines in audible and lipread speech intelligibility. Psychology and Aging, 36(4), 520–530. https://doi.org/10.1037/pag0000613

Dosher, B., & Lu, Z.-L. (2017). Visual perceptual learning and models. Annual Review of Vision Science, 3(1), 343–363. https://doi.org/10.1146/annurev-vision-102016-061249

Dryden, A., Allen, H. A., Henshaw, H., & Heinrich, A. (2017). The association between cognitive performance and speech-in-noise perception for adult listeners: A systematic literature review and meta-analysis. Trends in Hearing, 21, 2331216517744675. https://doi.org/10.1177/2331216517744675

Eberhardt, S. P., Auer, E. T., Jr., & Bernstein, L. E. (2014). Multisensory training can promote or impede visual perceptual learning of speech stimuli: Visual-tactile vs. visual-auditory training. Frontiers in Human Neuroscience, 8, 829. https://doi.org/10.3389/fnhum.2014.00829

Eberhardt, S. P., Bernstein, L. E., Demorest, M. E., & Goldstein, M. H., Jr. (1990). Speechreading sentences with single-channel vibratotactile presentation of voice fundamental frequency. The Journal of the Acoustical Society of America, 88(3), 1274–1285. https://doi.org/10.1121/1.399704

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech and Hearing Research, 12(2), 423–425. https://doi.org/10.1044/jshr.12.02.423

Ferguson, M. A., Henshaw, H., Clark, D. P., & Moore, D. R. (2014). Benefits of phoneme discrimination training in a randomized controlled trial of 50- to 74-year-olds with mild hearing loss. Ear and Hearing, 35(4), e110-e121. https://doi.org/10.1097/AUD.0000000000000202

Files, B. T., Auer, E. T., Jr., & Bernstein, L. E. (2013). The visual mismatch negativity elicited with visual speech stimuli. Frontiers in Human Neuroscience, 7, 371. https://doi.org/10.3389/fnhum.2013.00371

Files, B. T., Tian, B., Jiang, J., & Bernstein, L. E. (2015). Visual speech discrimination and identification of natural and synthetic consonant stimuli. Frontiers in Psychology, 6, 878. https://doi.org/10.3389/fpsyg.2015.00878

Fisher, C. G. (1968). Confusions among visually perceived consonants. Journal of Speech and Hearing Research, 11(4), 796–804. https://doi.org/10.1044/jshr.1104.796

Friston, K. (2005). A theory of cortical responses. Philosophical Transactions: Biological Sciences, 360(1545), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/10.1038/nrn2787

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 6(1), 110–125. https://doi.org/10.1037/0096-1523.6.1.110
Gesi, A. T., Massaro, D. W., & Cohen, M. M. (1992). Discovery and expository methods in teaching visual consonant and word identification. *Journal of Speech and Hearing Research, 35*(5), 1180–1188. https://doi.org/10.1044/jshr.3505.1180

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49*(1), 585–612. https://doi.org/10.1146/annurev.psych.49.1.585

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104*(4), 2438–2450. https://doi.org/10.1121/1.423751

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *The Journal of the Acoustical Society of America, 47*(2B), 613–617. https://doi.org/10.1121/1.1911936

Hall, D. A., Fussell, C., & Summerfield, A. Q. (2005). Reading fluent speech from talking faces: Typical brain networks and individual differences. *Journal of Cognitive Neuroscience, 17*(6), 939–953. https://doi.org/10.1162/0898929054021175

Hammer, R., Sloutsky, V., & Grill-Spector, K. (2015). Feature saliency and feedback information interactively impact visual category learning. *Frontiers in Psychology, 6*, 74. https://doi.org/10.3389/fpsyg.2015.00074

Hazen, V., & Rosen, S. (1991). Individual variability in the perception of cues to place contrasts in initial stops. *Journal of the Acoustical Society of America, 90*(4), 1317–138. https://doi.org/10.1121/1.411936

Henshaw, H., & Ferguson, M. A. (2018). Benefit from non-linear frequency compression hearing aids in a clinical setting: The effects of duration of experience and severity of high-frequency hearing loss. *International Journal of Audiology, 57*(5), 474–478. https://doi.org/10.1037/xhp0000410

Henshaw, H., & Ferguson, M. A. (2013). Efficacy of individual computer-based auditory training for people with hearing loss: A systematic review of the evidence. *PLOS ONE, 8*(5), Article e62836. https://doi.org/10.1371/journal.pone.0062836

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Ear and Hearing, 40*(2), 121–126. https://doi.org/10.1037/xhp0000410

Keating, P. A., Cho, T., Mattys, S., Bernstein, L. E., Chaney, B., Baroni, M., & Alwan, A. (2000). Articulation of word and sentence stress. *The Journal of the Acoustical Society of America, 108*(5), 2466. https://doi.org/10.1121/1.4743090

Keetels, M., Pecoraro, M., & Vroomen, J. (2015). Recalibration of auditory phonemes by lipread speech is ear-specific. *Cognition, 141*, 121–126. https://doi.org/10.1016/j.cognition.2015.04.019

Kollmeier, B., & Kiessling, J. (2018). Functionality of hearing aids: State-of-the-art and future model-based solutions. *International Journal of Audiology, 57*(Suppl 5), S3–S28. https://doi.org/10.1080/14992027.2016.1256504

Lalonde, K., & Werner, L. A. (2019). Perception of incongruent audiovisual English consonants. *PLOS ONE, 14*(4), Article e0213588. https://doi.org/10.1371/journal.pone.0213588

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research, 42*(3), 526–539. https://doi.org/10.1044/jslhr.4203.526

Lawrence, B. J., Jayakody, D. M. P., Henshaw, H., Ferguson, M. A., Eikelboom, R. H., Loftus, A. M., & Friedland, P. L. (2018). Auditory and cognitive training for cognition in adults with hearing loss: A systematic review and meta-analysis. *Trends in Hearing, 22*, 1–20. https://doi.org/10.1177/2331216518792096

Lecumberri, M. L. G., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication, 52*(11–12), 864–886. https://doi.org/10.1016/j.specom.2010.08.014

Lesner, S. A., Sandridge, S. A., & Kricos, P. B. (1987). Training influences on visual consonant and sentence recognition. *Ear and Hearing, 8*(5), 283–287. https://doi.org/10.1097/00004346-198710000-00005

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

Liu, J., Lu, Z. L., & Dosher, B. A. (2010). Augmented Hebbian reweighting: Interactions between feedback and training accuracy in perceptual learning. *Journal of Vision, 10*(10), 29. https://doi.org/10.1167/10.10.29

Liu, J., Lu, Z. L., & Dosher, B. A. (2012). Mixed training at high and low accuracy levels leads to perceptual learning without feedback. *Vision Research, 61*, 15–24. https://doi.org/10.1016/j.visres.2011.12.002
with moderate sensorineural hearing loss. *Ear and Hearing*, 51(1), 30–36. https://doi.org/10.1097/00003446-199401000-00007

Nahum, M., Nelken, I., & Ahissar, M. (2010). Stimulus uncertainty and perceptual learning: Similar principles govern auditory and visual learning. *Vision Research*, 50(4), 391–401. https://doi.org/10.1016/j.visres.2009.09.004

Nidiffer, A. R., Cao, C. Z., O’Sullivan, A., & Lalor, E. C. (2021). A linguistic representation in the visual system underlies successful lipreading. bioRxiv. https://doi.org/10.1101/2021.02.09.430299

Nittroer, S., Tarr, E., Wucinich, T., Moberly, A. C., & Lowenstein, J. H. (2015). Measuring the effects of spectral smearing and enhancement on speech recognition in noise for adults and children. *The Journal of the Acoustical Society of America*, 137(4), 2004–2014. https://doi.org/10.1121/1.4916203

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(5), 1209–1228. https://doi.org/10.1037/0278-7393.21.5.1209

Odegaard, B., & Shams, L. (2016). The brain’s tendency to bind audiovisual signals is stable but not general. *Psychological Science*, 27(4), 583–591. https://doi.org/10.1177/0956797616628860

Odegaard, B., Wozy, D. R., & Shams, L. (2017). A simple and efficient method to enhance audiovisual binding tendencies. *PeerJ*, 5, e3143. https://doi.org/10.7717/peerj.3143

Opoku-Baah, C., Schoenhaut, A. M., Vassall, S. G., Tovar, D. A., Ramachandran, R., & Wallace, M. T. (2021). Visual influences on auditory behavioral, neural, and perceptual processes: A review. *Journal of the Association for Research in Otolaryngology*, 22(4), 365–386. https://doi.org/10.1016/j.jaro.2021.02.0079

Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal hearing adult viewers. *Journal of Speech and Hearing Research*, 28(3), 381–393. https://doi.org/10.1044/jshr.2803.381

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, 68, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *The Journal of the Acoustical Society of America*, 55(2), 328–333. https://doi.org/10.1121/1.1914506

Raphael, L. J. (1971). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consomants in American English. *The Journal of the Acoustical Society of America*, 51(4B), 1296–1303. https://doi.org/10.1121/1.1912974

Rennig, J., Wegner-Clemens, K., & Beauchamp, M. (2020). Face viewing behavior predicts multisensory gain during speech perception. *Psychonomic Bulletin & Review*, 27(1), 70–77. https://doi.org/10.3758/s13423-019-01665-y

Rönnberg, J. (1995). Perceptual compensation in the deaf and blind: Myth or reality? In R. A. Dixon & L. Bäckman (Eds.), *Compensating for psychological deficits and declines* (pp. 251–274). Erlbaum.

Rönnberg, J., Holmer, E., & Rudner, M. (2019). Cognitive hearing science and ease of language understanding. *International Journal of Audiology*, 58(5), 247–261. https://doi.org/10.1080/14992027.2018.1551631

Rönnberg, J., Lunner, T., Zekveld, A., Sorqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience*, 7, 31. https://doi.org/10.3389/fnsys.2013.00031

Royle, L., Gaskins, C., Gordon-Salant, S., Goupell, M. J., & Anderson, S. (2019). Age effects on neural representation and
