Chapter 3
Toward a Model of Auditory-Visual Speech Intelligibility

Ken W. Grant and Joshua G. W. Bernstein

Abstract A significant proportion of speech communication occurs when speakers and listeners are within face-to-face proximity of one other. In noisy and reverberant environments with multiple sound sources, auditory-visual (AV) speech communication takes on increased importance because it offers the best chance for successful communication. This chapter reviews AV processing for speech understanding by normal-hearing individuals. Auditory, visual, and AV factors that influence intelligibility, such as the speech spectral regions that are most important for AV speech recognition, complementary and redundant auditory and visual speech information, AV integration efficiency, the time window for auditory (across spectrum) and AV (cross-modality) integration, and the modulation coherence between auditory and visual speech signals are each discussed. The knowledge gained from understanding the benefits and limitations of visual speech information as it applies to AV speech perception is used to propose a signal-based model of AV speech intelligibility. It is hoped that the development and refinement of quantitative models of AV speech intelligibility will increase our understanding of the multimodal processes that function every day to aid speech communication, as well guide advances in future generation hearing aids and cochlear implants for individuals with sensorineural hearing loss.

Keywords Articulation index · Auditory-visual coherence · Hearing loss · Modeling · Place of articulation · Spectrotemporal modulation index · Speech envelope · Speech intelligibility index · Speechreading · Speech transmission index · Temporal asynchrony · Temporal window of integration · Voicing

K. W. Grant (✉) · J. G. W. Bernstein
National Military Audiology and Speech Pathology Center, Walter Reed National Military Medical Center, Bethesda, MD, USA
e-mail: kenneth.w.grant.civ@mail.mil; joshua.g.bernstein.civ@mail.mil

A. K. C. Lee et al. (eds.), *Multisensory Processes*, Springer Handbook of Auditory Research 68, https://doi.org/10.1007/978-3-030-10461-0_3

3.1 Introduction

3.1.1 The Importance of Signal-Based Models of Speech Intelligibility

There can be little doubt of the importance of speech and language skills for cognitive and social development and for the communication of ideas, thoughts, and emotions. For the better part of a century, researchers have been working to develop models of speech perception and language processing, in large part due to the work at AT&T (Bell Laboratories) in the early 1900s. Driven by the development of the telephone and the need for high-quality speech transmission, the research team at Bell Laboratories developed a variety of methods for measuring speech intelligibility and user reactions to the phone. Among the many important discoveries stemming from this work was a characterization of how the signal-to-noise ratio (SNR), loudness, spectral balance, and distortion each impact speech intelligibility. Because of the expensive costs associated with test development and conducting laboratory and field experiments with human listeners, French and Steinberg (1947) and Fletcher and Gault (1950) began to work on methods for predicting the average speech quality of a given communication system as a means of testing new systems before they were put into the field. This work culminated in what became known as the articulation index (AI; American National Standards Institute [ANSI] 1969), which was designed to characterize a device, whether it be a phone, hearing aid, or any sound-transmitting system, based solely on the physical characteristics of the signal output and the environmental noise at the listener's ear.

3.1.2 The Overlooked Problem of Auditory-Visual Speech Intelligibility

Since its development, numerous extensions and simplifications of the AI or alternative metrics based on similar ideas have been proposed to predict speech intelligibility performance in different types of background noise (e.g., steady-state and modulated noise), reverberant environments, and for listeners with hearing impairment (speech intelligibility index [SII], ANSI 1997; speech transmission index [STI], Steeneken and Houtgast 2002). Despite the various iterations of these indices throughout the years, one of the most fundamental facts of human speech communication has been barely examined, namely, that human communication involves auditory-visual (AV) face-to-face input and not just auditory input. It is estimated that well over half of active speech communication takes place in contexts where visual speech cues are available to the listener (Walden et al. 2004). Yet, the prediction of intelligibility for AV speech inputs is woefully underdeveloped. The AI and SII ANSI standards did include a nod to AV speech recognition, but visual cues were treated simply as an additive factor to the basic auditory predictions and failed

to understand the intricate manner in which auditory and visual speech cues interact. The goals of this chapter are to illuminate the factors that would necessarily be an important part of any AV speech-intelligibility model and to suggest solutions that are consistent with the original goals of the AI. In addition to being able to accurately predict AV speech intelligibility under a wide variety of noise and reverberation conditions, a practical model should be based on physical measurements of the signal and environment alone to allow for the evaluation of potential benefits of new hearing technologies and algorithms without relying on exhaustive human-subjects testing. (Ideally, any auditory or AV model of speech intelligibility would also consider individual differences in dimensions such as hearing acuity, visual acuity, and cognitive ability; however, accounting for individual differences falls outside the scope of this chapter.) In delineating the factors involved in the development of such a model, this chapter will revisit some of the same issues that had to be addressed during the development of the original auditory-only (A-only) AI. This will include (1) impact of noise and distortion, (2) spectral balance or frequency weighting, (3) integration across spectral channels and across modality, and (4) synchrony between auditory and visual signals.

With few exceptions, listeners are able to improve their speech-recognition performance by combining visual cues (from lipreading; also known as speechreading) and audition (e.g., Sumby and Pollack 1954; Grant et al. 1998). Benefits due to speechreading, especially in reverberant or noisy environments, can be quite substantial for most listeners, often allowing near-perfect comprehension of otherwise unintelligible speech (Grant et al. 1985; Summerfield 1992). Understanding how these large benefits come about is critical because the speech cues that must be relayed to maximize speech understanding in adverse situations are likely to be dramatically different when the listener has access to visual (speechread) cues in addition to acoustic speech information. As discussed, this is the case when considering normal-hearing (NH) listeners in adverse noisy listening environments, hearing-impaired (HI) listeners, or signal-processing strategies for hearing aids and advanced auditory prosthetics such as cochlear implants.

Consider the following scenario (see Fig. 3.1). A speech signal composed of both visual and acoustic information is presented. The listener-observer extracts signal-related segmental (i.e., phonemes and syllables) and suprasegmental (i.e., words and phrases) cues from each modality, integrates these cues, and applies top-down semantic and syntactic constraints in an effort to interpret the message before making a response. The basic components—bottom-up signal-related cue extraction, integration, and top-down linguistic processes—are common to most speech-perception theories (e.g., Liberman et al. 1967; Studdert-Kennedy 1974). The major distinction drawn here from A-only theories of speech perception is that in an AV communication environment, cues from the visual modality must be considered, and the integration of auditory and visual cues, both within and across modalities, must be evaluated (Massaro 1987). From this perspective, consider an individual whose AV recognition of words and sentences is less than perfect. To evaluate the exact nature of the communication problem, it is necessary to determine whether the deficit is due to poor reception of auditory or visual cues, difficulty in integrating

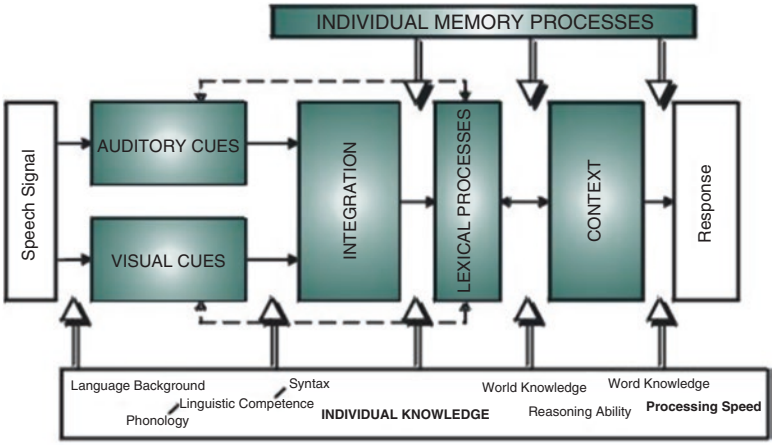

Fig. 3.1 A schematic of the predominant sources of individual variability in auditory-only (A-only) and auditory-visual (AV) speech processing. Processing starts with the common assumption of sensory independence during the early stages of processing. The integration module as a potential source of individual variability uses a model of optimal AV processing

auditory and visual cues, difficulty in applying linguistic and contextual constraints, cognitive limitations such as reduced working-memory capacity or reduced attention, or a combination of these factors. If the problem is determined to be primarily difficulty in receiving auditory or visual cues, signal-processing strategies to enhance the relevant cues and improve the SNR may be used. If, on the other hand, the problem is determined to be difficulty in integrating auditory and visual cues or difficulty in applying top-down language-processing rules, then training and practice techniques may be the better rehabilitation strategy. Simply knowing the individual's AV sentence- or word-recognition performance is not sufficient for determining a plan for rehabilitation.

Based on the simple framework displayed in Fig. 3.1, three questions must be addressed in order to predict speech intelligibility. (1) What are the most important cues for AV speech recognition that can be extracted from acoustic and visual speech signals? (2) How should one measure an individual's ability to integrate auditory and visual cues separate and apart from their ability to recognize syllables, words, and sentences? (3) What are the most important non-signal-related "top-down" processes that contribute to individual variability in AV speech recognition? Because the top-down influences on speech recognition are quite influential, early models of speech intelligibility and most models of AV speech intelligibility and integration limit the types of speech materials used to include mostly nonsense syllables (French and Steinberg 1947; Fletcher 1953). By imposing this limitation on the types of speech signals considered, the focus of the model becomes "bottom-up" and highly dependent on the signal, room, and any equipment (e.g., radio, phone) that resides in the transmission path between the speaker and listener.

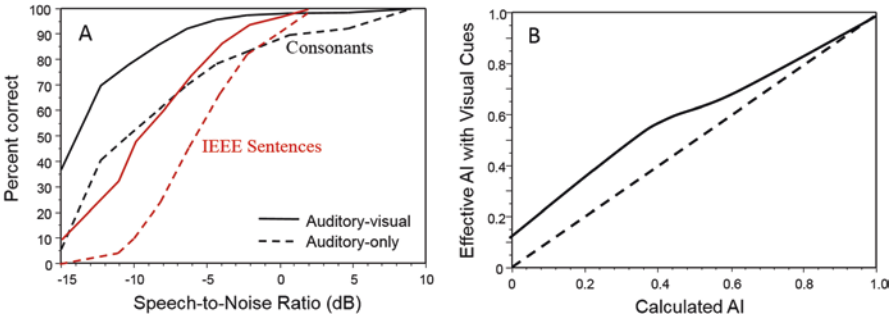

Fig. 3.2 (**A**) Impact of noise on A-only (*dashed lines*) and AV (*solid lines*) speech perception for sentence recognition (*red curves*) and consonant recognition (*black curves*). (**B**) Relationship between calculated articulation index (AI) and "effective" AI when auditory cues are combined with speechreading (from American National Standards Institute [ANSI] 1969)

Figure 3.2A shows a typical outcome for A-only (*dashed lines*) and AV (*solid lines*) speech recognition of low-context sentences (*red curves*) and consonants (*black curves*) for NH young adults (after Grant and Braida 1991; Grant and Walden 1996). For both sets of speech materials, performance was substantially better in the AV condition. At a SNR of −15 dB, the auditory signal was just audible, with performance at zero percent correct for sentences and at chance level for consonants (i.e., 1/18 response options). AV keyword recognition scores at a SNR of −15 dB were roughly 10% correct for sentences. For consonant materials, however, the AV scores at −15 dB SNR were near 40% correct. As will be discussed below in Sect. 3.1.3, this can be explained by the fact that although speechreading alone can barely support word recognition, it can convey very accurate information about certain aspects of speech.

The original ANSI (1969) standard for calculating the AI included an overly simplistic graphical solution to predict the presumed benefit to intelligibility when visual speech cues are present (Fig. 3.2B). In the revised version of the AI known as the SII (ANSI 1997), the effective benefit of visual cues was formalized by a simple two-part equation, essentially mimicking the curve shown in the ANSI (1969) standard. An unmistakable conclusion one can draw from Fig. 3.2B is that the addition of visual cues to speech intelligibility was treated as an effective addition to the AI and that the same AV prediction would be made for a given level of A-only performance regardless of the particular spectral characteristics of the speech signal and noise. In other words, the importance of different spectral regions for A-only intelligibility was assumed to be the same for AV intelligibility.

We now know this assumption to be incorrect. HI listeners show dramatic benefits from speechreading in cases with very little residual auditory function (Erber 1972; Drullman and Smoorenburg 1997). Studies of NH listeners have allowed us to understand this phenomenon. When speechreading is combined with low-frequency, low-intelligibility auditory speech cues, the resulting benefits are enormous. Grant et al. (1985) found that even presenting a sparse acoustic representation of the speech

cues located at these low frequencies was sufficient to generate large speechreading benefits on the order of 50 or more percentage points. Adding low-frequency speech signals dramatically sped up the ability to track connected discourse (by repeating back verbatim text read aloud), from 41 words per minute (wpm) for speechreading alone up to 86 wpm for AV speech (tracking rates for full bandwidth speech were 108 wpm). Similarly, Rosen et al. (1981) showed that presenting only the acoustic voice-pitch information provided an 83% improvement in the rate of discourse tracking over speechreading alone. These extremely large increases in the ability to track AV speech when the low-frequency acoustic signals produced zero percent intelligibility by themselves indicate that AV intelligibility does not completely depend on A-only intelligibility as suggested by the AI and SII. Instead, an accurate prediction of AV intelligibility requires an understanding of the information provided by the auditory and visual signals. In particular, Grant and Walden (1996) showed that the addition of visual cues enhances auditory speech perception for low-frequency stimuli much more than for high-frequency stimuli. As will be discussed in Sect. 3.1.3, this is because the visual signal and low-frequency auditory signals provide complementary information. The visual signal facilitates the differentiation of visible speech features generated at the lips (e.g., /ba/ vs. /ga/), whereas the low-frequency auditory signal facilitates the differentiation of invisible speech features generated in the back of the throat or at the larynx (i.e., /ba/ vs. /pa/).

In cases where A-only speech intelligibility is impacted by hearing loss and not just environmental conditions, the importance of speechreading in everyday communication settings increases. Furthermore, when auditory and visual speech cues are integrated successfully, the improvement to speech intelligibility can be so large that the benefit from speechreading can even outweigh the benefit from a hearing aid. Walden et al. (2001) reported consonant-recognition data from 25 adults (mean age 66 years) with an acquired moderate-to-severe high-frequency sensorineural hearing loss. The benefit of visual cues compared with unaided listening was roughly 40 percentage points, whereas the benefit of amplification was only 30 percentage points. (Although this experiment was conducted with older hearing-aid technology, the benefits of amplification for speech understanding in quiet are mostly unaffected by newer technological advances.) A small additional benefit was observed when hearing aids were combined with speechreading, although ceiling effects likely obscured some of the benefits from combining amplification and speechreading. The small difference between aided and unaided AV scores could conceivably contribute to the listener's notion that the hearing aids were not that beneficial under typical AV conditions. In another example where the presence of visual speech might obscure the benefits of newer hearing-aid technologies, directional microphones for improving the SNR are a key feature of almost all modern hearing aids. When evaluated without visual cues, this feature can provide a substantial improvement in SNR (3–5 dB in indoor environments and 4–8 dB in outdoor environments; Killion et al. 1998). However, when evaluated with visual cues, the perceived and objective benefit of directional microphones can be greatly reduced (Wu and Bentler 2010). Thus, even if an advantageous hearing-aid feature is developed that proves to be very useful in an A-only listening situation, it is not guaranteed to be equally beneficial (or even noticed) in an AV listening situation.

In summary, the studies of Grant et al. (1985), Grant and Walden (1996), and Walden et al. (2001) demonstrate two important concepts. First, the advantages for speech understanding provided by integrating auditory and visual speech cues are determined by a complex interaction among auditory and visual speech information as well as several important top-down influences. This means that AV speech-reception performance cannot be predicted from A-only speech-reception performance without an understanding of the information relayed by each modality and some assessment of information redundancy and complementarity. Second, the effects of hearing loss and hearing aids might be very different under AV and A-only conditions. The most commonly used hearing-aid fitting algorithms are based on maximizing model-predicted A-only speech intelligibility (e.g., Byrne et al. 2001). The fact that AV speech perception is likely to depend on hearing loss and hearing-aid features differently than A-only speech perception highlights the need for an AV model of speech intelligibility.

Because of the importance of visual cues for speech communication and the fact that speechreading and auditory cues interact in a nonadditive manner, studies measuring the contribution of these cues to speech perception and theories of AV speech perception have become more common in the literature (see Summerfield 1987; Massaro 1998 for reviews). Furthermore, despite the obvious importance of speech communication for maintaining the health and fitness of elderly persons, little is known about the combined effects of hearing loss, visual acuity, and aging on AV speech recognition, making the task of developing an AV version of the AI that much more difficult. However, for the purposes of this chapter and in the spirit of the original AI, the first step of accurately modeling AV speech recognition for a NH population is the current focus, leaving aside for now the more complex questions related to sensory impairment and individual variability (hearing loss, aging, visual acuity, and cognitive decline).

3.1.3 Speech-Feature Complementarity and the Relative Importance of Different Spectral Regions

How can an acoustic signal that generates near zero intelligibility on its own so dramatically improve speechreading performance? An important clue to understanding this synergy comes from research that has carefully analyzed the pattern of particular errors that listeners make in speech-recognition tests (Grant and Walden 1996). These analyses show that some of most reliable information relayed by speechreading are surface features of the lips and tip of the tongue that help to differentiate between certain consonants. For example, by speechreading alone, it is very easy to tell the difference between /bɑ/, /gɑ/, and /dɑ/, even though these tokens would often be confused in the case of A-only speech processing in a noisy situation or by listeners with hearing loss. In contrast, speechreading provides very little information regarding speech contrasts generated at the larynx. For example, visual representations of /bɑ/, /pɑ/, and/mɑ/ are often confused with one another. Although not usually enough to support high levels of intelligibility, being able to accurately

recognize these visual categories of contrast greatly reduces the number of possible choices when making a response. When combined with the right kind of complementary acoustic information, the integration of auditory and visual speech cues can lead to very high levels of speech intelligibility.

To illustrate the complementary nature of auditory and visual speech cues, it is useful to discuss the relative contributions of the speech signals in terms of their articulatory and phonetic distinctive features (voicing [e.g., /bɑ/ vs. /pɑ/], manner of articulation [e.g., /bɑ/ vs. /mɑ/] and place of articulation [e.g., /bɑ/vs. /gɑ/]). Briefly, *place of articulation* refers to the location within the vocal tract where the airflow has been maximally restricted. For example, the primary place of constriction for the consonant /mɑ/ is the lips. Place-of-articulation cues are often clearly visible in terms of lip and tongue-tip position. Acoustically, these dynamic high-frequency speech cues associated with the second and third formant transitions are considered to be fragile and easily corrupted by noise or hearing loss (Kewley-Port 1983; Reetz and Jongman 2011). *Voicing* cues mainly reflect the presence or absence of waveform periodicity or vocal-fold vibration. Taking place in the larynx, these cues are not visibly apparent. Acoustically, voicing is well represented in the low frequencies of speech and is marked by attributes such as voice-onset time and the trajectory of the first formant immediately following the consonant release (Reetz and Jongman 2011). *Manner of articulation* refers to the way the speech articulators interact when producing speech. For example, for the consonant /s/, the tip of the tongue forms a narrow constriction with the alveolar ridge (gum line) just behind the teeth. The result of this constriction is a turbulent airflow that serves as the primary source of the sound, making /s/ a fricative. These three broad phonetic and articulatory features are not orthogonal, although each sound in English can be uniquely identified by a combination of place, manner, and voicing (e.g., /bɑ/ is classified as a voiced, bilabial, plosive; /sɑ/ is classified as a voiceless, alveolar, fricative).

Figure 3.3 illustrates how auditory and visual information interact across these three types of consonant feature. Each panel shows the percentage correct in identifying a particular consonant feature under A-only and AV conditions (Grant and Walden 1996). Figure 3.3C shows that place-of-articulation information is readily available to the speechreader, is not affected by noise, and does not need auditory place cues to reach ceiling performance. In contrast, voicing information (Fig. 3.3A) is determined entirely by auditory cues with very little contribution from the visual speech signal. Figure 3.3B shows the results for manner of articulation and, at first glance, suggests that visual information is helpful for making consonantal manner determinations and combines with auditory cues as they become available with improving SNR. However, further analysis (not shown) suggests that this is due to the high degree of redundancy between place and manner cues for consonant identification. In other words, the score observed for manner information by speechreading alone is what one would predict by chance given 100% correct transmission-of-place information. Thus, for these consonant materials, speechreading contributes almost exclusively to the reception of place information.

Grant and Walden (1996) also provided insight into how the complementarity of speech features (Fig. 3.3) translates into a complex interaction between speechreading

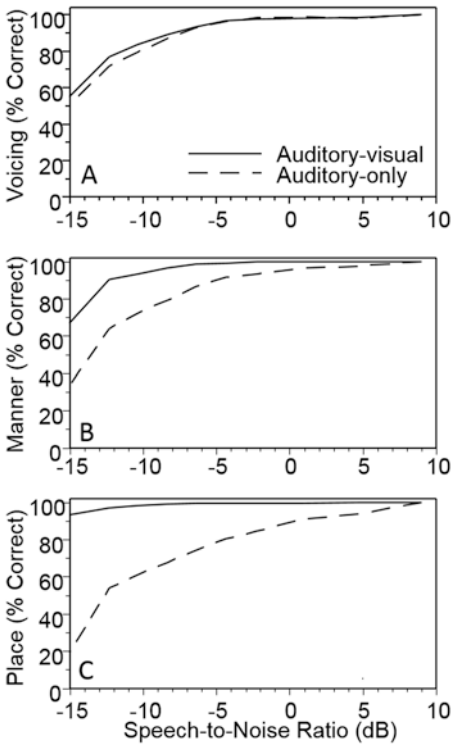

Fig. 3.3 A-only and AV feature transmission for consonant identification (Grant and Walden 1996). The information contained in the visual signal is derived by comparing A-only and AV performance for each feature. Visual cues contribute almost zero information regarding voicing (**A**), some manner information (**B**), and near perfect place-of-articulation information (**C**)

benefit and the spectral content of the speech signal. The AI makes the basic assumption that better A-only speech-reception performance will also result in better AV performance (Fig. 3.2B). In contrast, when Grant and Walden examined the speechreading benefit for filtered bands of speech, they found that the AV speech scores did not increase monotonically with A-only performance. Instead, speechreading benefit varied substantially depending on the filter bandwidth and center frequency, even for frequency bands that generated equal A-only performance. Twelve bandpass-filter conditions were chosen to carefully control the A-only AI prediction while varying the bandwidth and center frequency of the filter. Figure 3.4A shows the results, with the A-only conditions (solid bars) arranged in ascending order based on percentage- correct consonant-identification scores. The AV speech-reception scores were only weakly correlated with A-only performance, demonstrating clear nonmonotonicity between A-only and AV speech recognition. The relationship between AV benefit and spectral region is clearly exemplified in the comparison between filter conditions 1 and 6. Whereas filter condition 6 (containing high-frequency speech information) yielded a substantially higher A-only speech score, AV performance was substantially better in condition 1 (containing only low-frequency speech information). This pattern was observed repeatedly across filter-band conditions (e.g., compare conditions 7 and 9 and conditions 10 and 12). (It should be noted that this same pattern of results holds whether the difference between AV and

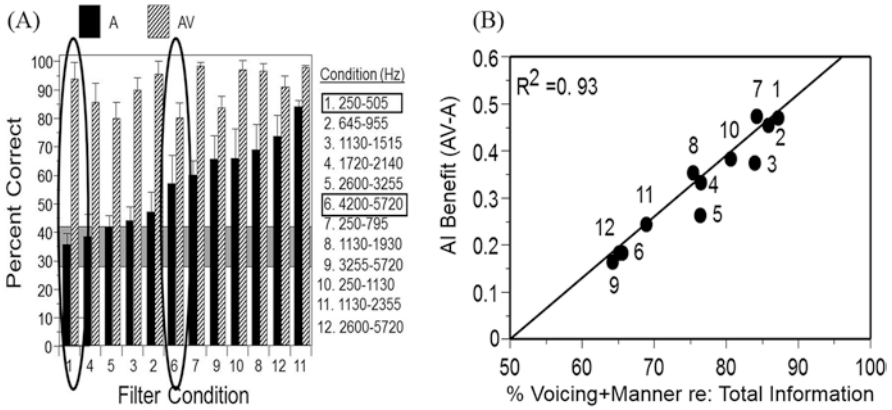

Fig. 3.4 (**A**) Consonant recognition scores for A-only and AV filtered speech. *Horizontal gray band* between 30 and 40% correct reflects the range of speechreading-only scores. Ellipses highlight two conditions (filter bands 1 and 6) representing a narrow low-frequency and a high-frequency band, respectively. Note that although A-only performance is significantly greater for band 6 than for band 1, the AV score for high-frequency band 6 is much less than that for band 1, demonstrating nonmonotonicity between A-only and AV performance. (**B**) Visual benefit as predicted by the proportion of voicing plus manner-of-articulation information relative to the total amount of transmitted information for the 12 bandpass-filtered conditions tested. The greatest AV benefit occurs for filtered speech with a high concentration of low-frequency energy and a high relative transmission rate for voicing and manner information. From Grant and Walden (1996)

A-only speech scores are measured in terms of percentage correct or as relative benefit, taking into account how close to ceiling performance the A-only score might be; Sumby and Pollack 1954).

The results in Fig. 3.4A show that AV intelligibility was greater when the audible acoustic speech spectrum was dominated by low-frequency energy than when it was dominated by high-frequency energy. This suggests that the frequencies that are most important for speech understanding are very different under A-only conditions (mid-to-high frequencies; ANSI 1969) than under AV conditions (low frequencies). To investigate why low-frequency auditory information is so highly complementary with visual speechreading cues, Grant and Walden (1996) examined the relationship between an information analysis of consonant features (Miller and Nicely 1955) and the frequency dependence of the speechreading benefit (Fig. 3.4B). This analysis clearly showed that the magnitude of the AI benefit due to the addition of visual cues was strongly correlated with the amount of voicing and manner of information transmitted for a given frequency band. Low-frequency conditions (e.g., bands 1, 2, 3, 7, and 10) transmitted a great deal of voicing and manner information relative to the total amount of information contained in each band and generated the largest AI benefit. The reverse was true for high-frequency conditions (e.g., bands 6, 9, and 12). Comparable analyses of the visual-only (V-only) condition confirmed that the low-frequency auditory bands were essentially complementary with speechreading while the high-frequency bands were mostly redundant with speechreading. In other words,

the reason that visual speech cues provided such a large benefit when the auditory signal was limited to low frequencies is because voicing (and manner) information available at low frequencies was highly complementary to the place information provided by the visual signal. In contrast, the visual signal provided much less additional benefit for high-frequency stimuli because both modalities contributed largely redundant place information.

3.1.4 Auditory-Visual Integration Efficiency

The integration of the auditory and visual modalities of speech information requires a neural process that combines the two inputs and hence could be susceptible to individual differences in integration efficiency (see Fig. 3.1). In fact, it is often assumed that if a particular stimulus condition demonstrates a large visual benefit to speech intelligibility, then the listener must have been able to integrate auditory and visual information with a high degree of efficiency (Sommers et al. 2005). However, as just demonstrated, the processes involved in integrating auditory and visual information efficiently and the amount of AV benefit obtained compared to A-only or V-only intelligibility are distinctly different processes. As shown in Fig. 3.4, the amount of AV benefit observed is much more closely related to the spectral region of the acoustic speech signal than to the A-only or V-only recognition scores. Thus, the fact that one acoustic condition shows a much larger visual benefit than another could be because it provides access to very different auditory information and not necessarily because there is a problem integrating information across modalities. Stated another way, the fact that filter-band condition 6 demonstrated far less benefit than filter-band condition 1 (Fig. 3.4A) does not mean that AV integration was less efficient for filter-band 6. The question of integration efficiency can be specifically examined using a class of models of AV integration for consonant identification developed by Massaro (1987) and Braida (1991). These models take as input confusion matrices that describe the speech information contained in separate auditory and visual speech signals (or for separate frequency bands of auditory speech signals). They then make an AV prediction based on the mutual information contained in the V-only and A-only conditions. Grant et al. (2007) applied the modeling approach of Braida (1991), defining integration efficiency in terms of the ratio between the model prediction and the actual AV performance (or performance for combinations of auditory frequency bands). NH listeners were found to have nearly perfect integration efficiency both within and across modalities. HI listeners were found to have slightly reduced efficiency (although not significantly so) for combining auditory and visual speech information but significantly reduced efficiency for combining auditory speech information across frequency bands. Similarly, Tye-Murray et al. (2007) found that HI adults do not exhibit a reduced ability to integrate auditory and visual speech information relative to their age-matched, NH counterparts. Thus, HI listeners demonstrate greater difficulty integrating acoustic bands across the spectrum than they do integrating auditory and visual cues.

3.1.5 Auditory-Visual Asynchrony

Although the AV integration models of Massaro (1987) and Braida (1991) (and more generally of Alais and Burr, Chap. 2) can successfully account for the role of feature complementarity and redundancy in predicting AV speech intelligibility using probabilistic approaches such as fuzzy logic, multidimensional scaling, or maximum likelihood estimation, they each share an important shortcoming that prevents their wider application in the tradition of AI, SII, or STI models. To apply any of these models to the problem of AV integration, the uncertainty contributed by each separate modality regarding the identity of a given speech token or speech feature must be determined. In tests of nonsense-syllable identification, confusion matrices for A-only and V-only (at a minimum) must be obtained before any predictions of bimodal processing can take place (Massaro 1987; Braida 1991). Because these models as applied to speech identification require an abstraction of the auditory and visual speech information to phoneme labels before they are integrated, they cannot achieve what the AI methodology can accomplish by making speech-intelligibility predictions based on the physical properties of the speech signals alone.

Some clues for how one might accomplish the goal of a signal-based prediction of AV speech perception come from a number of studies that have examined how the temporal relationship between auditory and visual speech signals affects AV integration (see Fig. 3.5A). Studies have shown that AV integration does not require precise temporal alignment between A-only and V-only stimuli (e.g., McGrath and Summerfield 1985; Massaro et al. 1996). However, these studies also demonstrated

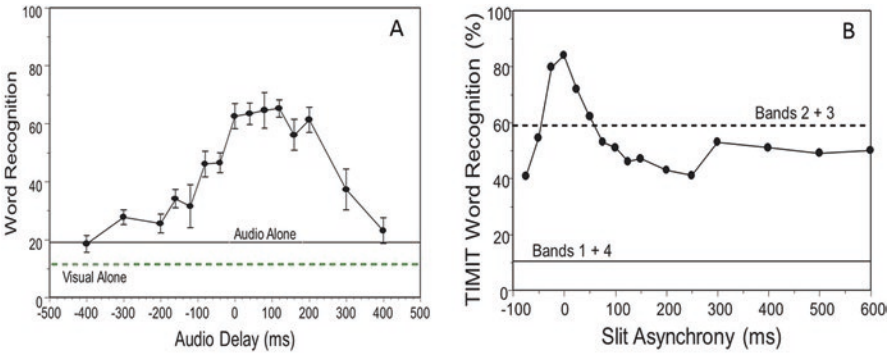

Fig. 3.5 (**A**) Average AV keyword intelligibility (low-context IEEE sentences) as a function of AV asynchrony. There is a substantial plateau region between approximately −50 ms (audio leading) to +200 ms (audio lagging) where intelligibility scores are high relative to the A-alone (*horizontal solid line*) or V-alone (*horizontal dashed line*) conditions. Error bars are ±1 SD. (**B**) Average A-only sentence intelligibility (Texas Instruments/Massachusetts Institute of Technology [TIMIT] sentences; Garofolo et al. 1990, 1993) for synchronous and asynchronous presentations of one-third octave, widely spaced auditory spectral slits. Unlike the AV condition, peak word-recognition performance in the A-only condition occurs when the different bandpass-filtered signals are presented synchronously and intelligibility falls off precipitously when any asynchrony is introduced across the spectral bands. From Grant et al. (2004)

that the temporal windows of integration (TWI) over which AV interactions can successfully occur are very asymmetric, with much greater tolerance found for visual-leading than for visual-lagging conditions. For naturally produced "congruent" speech, where the speaker's articulations and speech sounds are matched, auditory-lagging misalignments of up to 200 ms are easily tolerated, whereas visual-lagging misalignments greater than 20 ms lead to a breakdown in AV integration (Grant et al. 2004; Shahin et al. 2017). The asymmetry of the TWI favoring audio delays is consistent with the idea that for most speech utterances, the movement of the mouth begins before any sound is emitted. It has also been suggested that because visual speech information is available to the listener before the acoustic speech signal, it has the potential to facilitate language processing (e.g., lexical access) by allowing initial lexical pruning to proceed before any speech is heard (van Wassenhove et al. 2005, 2007). The fact that AV integration takes place over limited and multiple time windows suggests that bimodal speech processing is based on neural computations occurring at an earlier stage than a speech feature-based analysis.

In contrast to the long asymmetric temporal windows associated with AV integration, the TWI for combining information across acoustic frequency bands is both symmetric and narrow (see Fig. 3.5B). One interpretation of these data is that there are multiple time intervals over which speech is decoded in the auditory system. These include short-range analysis windows (1–40 ms), possibly reflecting various aspects of phonetic detail at the articulatory feature level (e.g., voicing); midrange analysis windows (40–120 ms), possibly reflecting segmental processing; and long-range analysis windows (beyond 120 ms), possibly reflecting the importance of prosodic cues, such as stress accent and syllable number, in the perception of running speech. The differences observed for cross-spectral (within modality) and cross-modal integration are important considerations for models of intelligibility as they highlight the different timescales associated with processing fine structure (formant transitions), syllabicity, and intonation. The different time frames may also implicate cortical asymmetries whereby left auditory areas process primarily short temporal integration windows while the right hemisphere processes information from long integration windows (Poeppel 2003). Yet the fact that the auditory and visual signals must be at least somewhat temporally coherent (Fig. 3.5A) suggests that a model of AV speech perception based on the coherence of auditory and visual signals might better represent the underlying process of AV integration than a feature-based or intelligibility-based approach.

3.1.6 Perception of Auditory-Visual Coherence and the Enhancement of the Auditory Speech Envelope

Another clue for how the auditory and visual speech signals might temporally interact comes from a set of speech-detection experiments conducted by Grant and Seitz (2000) and Grant (2001). The goal of these experiments was to determine whether movements of the lips perceived during speechreading could be used to improve the

masked detection thresholds of congruent auditory signals. The basic idea used a variant of the comodulation masking-release paradigm (Hall et al. 1984), but in this coherence-protection paradigm (Gordon 1997, 2000), the audio speech target and visible movements of the lips were comodulated while the masker (e.g., speech-shaped noise) was uncorrelated with the target speech signal. The fact that the movements of the lips were coherent with the audio speech envelopes should have helped to protect the target speech from being masked and therefore improve detection thresholds.

From a strictly psychophysical perspective, it is reasonable to assume that the correlation between lip movements and acoustic envelope would be useful in detecting speech in noise and, further, that the greatest synergistic effects would be seen for sentences with the highest correlations. This is exactly what was found in studies by Grant and Seitz (2000) and Grant (2001). These studies showed a significant masking release for detecting spoken sentences (1–3 dB depending on the particular sentence) when simultaneous and congruent visual speech information was provided along with the wideband acoustic speech signal (Fig. 3.6, AV_{WB}). Incongruent speech (not shown) had no effect and resulted in the same threshold as the A-only condition. Finally, knowing prior to each trial (by orthography; AV_O) the content of the specific sentence to be detected had a mild positive influence (roughly 0.5 dB masking release) and was independent of which particular sentence was presented.

Critically, Grant (2001) showed that the degree of AV masking protection was related to the degree to which the auditory and visual signal envelopes were correlated. Figure 3.7 shows the time-intensity waveform, amplitude envelopes, and area of mouth opening for the sentence "Watch the log float in the wide river" (similar relationships can be seen for almost any AV sentence with only minor variations in the results). The traces in Fig. 3.7A represent the envelope extracted from wideband (WB) speech and from the speech filtered into three different spectral bands repre-

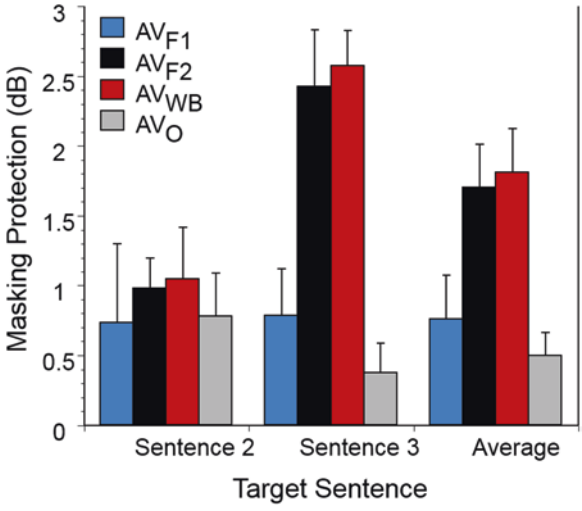

Fig. 3.6 Difference in A-only and AV masked detection thresholds (masking protection) for spoken filtered sentences (Grant 2001). $AV_{F1,}$ AV visual presentation of speech filtered between 100 and 800 Hz; AV_{F2}, AV presentation of speech filtered between 800 and 2200 Hz; AV_{WB}, AV presentation of wideband speech (100–8500 Hz); AV_O, auditory presentation of wideband speech preceded by visual orthography. Error bars show +1 SD

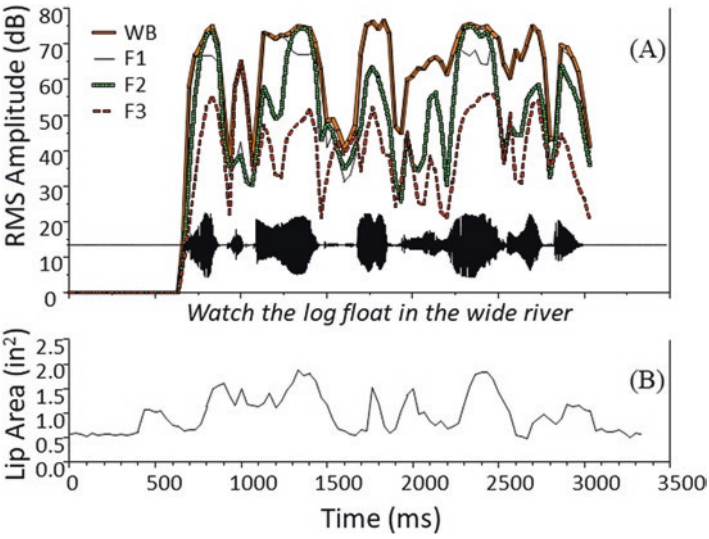

Fig. 3.7 (**A**) Waveform and amplitude envelopes extracted from wideband (WB) speech and from bandpass-filtered speech with filters centered at the F1 (100–800 Hz), F2 (800–2200 Hz), and F3 (2200–8500 Hz) formant regions. RMS, root-mean-square. (**B**) Amplitude envelope of the kinematic lip movements over time during speech production. The correlation between acoustic envelope and visual movement (area of mouth opening) was greatest for the envelope in the F2 region (0.65) and weakest in the F1 region (0.49). From Grant (2001)

senting the first (F1), second (F2), and third (F3) speech formants. These envelopes were clearly comodulated with the area of mouth opening extracted from the video image (Fig. 3.7B). However, the degree of correlation was largest for the acoustic envelopes derived from the higher frequency regions (F2 and F3) than for the F1 envelope. Grant (2001) found that WB speech or speech filtered into the F2 (800–2200 Hz) region also produced substantially more masking protection (about 2 dB on average) than for speech filtered into the F1 (100–800 Hz) region (less than 1 dB; Fig. 3.6, AV_{F2}, AV_{F1}). Thus, as long as the speech signal contained energy in the F2 region associated with place of articulation, the addition of synchronized visual information from the face of the speaker provided significant masking protection and lower detection thresholds. Overall, these results showed that listeners used the visible modulations of the lips and jaw during speechreading to make auditory detection easier by informing them about the probable spectrotemporal structure of a near-threshold acoustic speech signal, especially with peak energy in the F2 frequency range.

The temporal coherence of the acoustic and visual signals and the fact that the brain can make use of this temporal coherence to more easily detect the acoustic signal offer the possibility of analyzing the acoustic and visual signals within a single common mechanism of time-intensity envelope processing (see Lee, Maddox, and Bizley, Chap. 4). The fact that the modulation envelopes for speech of mid- to high-frequency auditory channels and the slowly time-varying visual kinematics of

the speaker's face (e.g., area of mouth opening over time) are strongly correlated with one another provides a mechanism for combining the auditory and visual inputs directly at the physical signal level without requiring lengthy and costly behavioral experimentation. Section 3.2 describes efforts toward the development of a signal-based model of AV speech perception that makes predictions based on (1) the coherence between the auditory and visual signals over long temporal windows of integration, (2) greater AV benefit relative to A-only speech recognition at poorer SNRs, and (3) greater correlation between visual kinematics and the acoustic envelopes in the higher speech frequencies.

3.2 Modeling Auditory-Visual Speech Intelligibility

A model of AV speech perception based on the temporal coherence of the auditory and visual modalities necessarily requires an analysis of the temporal modulations of speech across the frequency spectrum. In this regard, the model would share many characteristics of the STI (Steeneken and Houtgast 2002), a model that takes into account the degree of modulation degradation as a result of noise, reverberation, or hearing loss. By considering the dynamics of the visual speech signal as additional modulation channels that can be used to reduce some of the deleterious effects of noise and reverberation, this approach can be easily expanded to include the influence of speechreading on speech intelligibility.

Grant et al. (2008, 2013) described a signal-based AV speech-intelligibility model that considered both auditory and visual dynamic inputs, combining them at the level of the speech envelopes to generate a prediction of AV speech intelligibility in noise. The basic premise was that the brain can use the visual input signals to help reconstruct the temporal modulations inherent in the "clean" auditory signal (minus noise or reverberation) based on *a priori* knowledge of the relationship between facial kinematics and the temporal envelopes of the audio speech signal. This approach was inspired by the engineering applications of Girin et al. (2001) and Berthommier (2004) showing that a video signal of the talker's face could be used to enhance a noise-corrupted audio speech signal.

Grant et al. (2008, 2013) used a biologically inspired auditory spectrotemporal modulation index (STMI) model (Elhilali et al. 2003) to make A-only and AV speech-intelligibility predictions. Like the STI, the STMI bases its predictions on analysis of the critical modulations present in the speech signal. However, the STMI includes an additional dimension, spectral modulation, which is critical to the prediction of the effects of spectral smearing caused, for example, by the reduced frequency selectivity associated with hearing loss (Bernstein et al. 2013). The model (Fig. 3.8) consisted of three main stages: (1) a "peripheral" stage that processed the acoustic waveform into frequency bands and derived the envelope in each band, (2) a "cortical" stage that processed the resulting envelopes to derive the modulation spectra, and (3) an "evaluation" phase that compared the resulting spectrotemporal modulation profile of speech presented in noise with the profile associated with

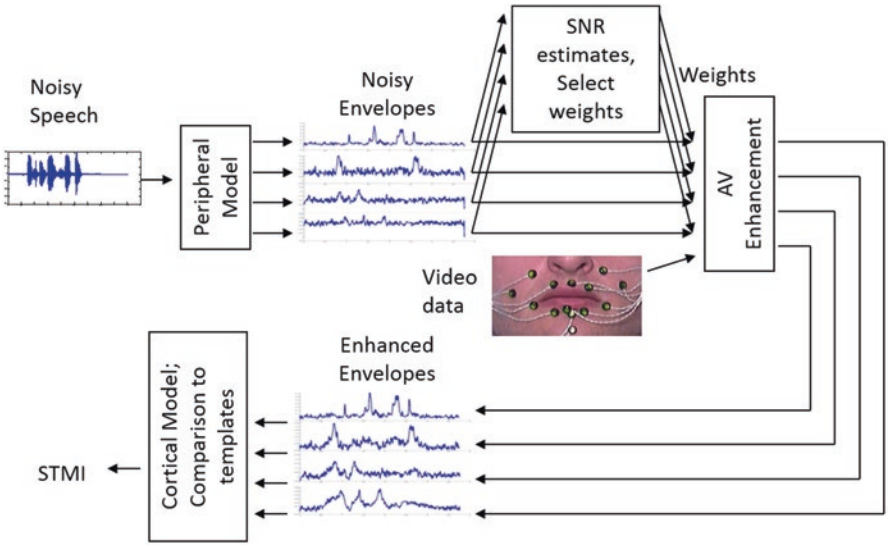

Fig. 3.8 Schematic of the expanded AV-spectrotemporal modulation index (STMI) model showing the inclusion of visual speech-movement envelopes to enhance the outputs of each auditory channel. The enhanced AV envelope channels were then processed by the cortical model and compared with "clean" speech templates to make the final AV speech-intelligibility estimate. *SNR* signal-to-noise ratio

clean speech ("comparison to speech templates" in Fig. 3.8). To extend the STMI model to include visual modulation channels, the model also included an "AV enhancement" component that cleaned up the noisy acoustic speech envelopes based on *a priori* knowledge about the relationship between the auditory and visual stimuli.

An example of the output of the peripheral stage of the model for an acoustic speech signal presented in speech-shaped noise is shown in Fig. 3.9A. Each individual curve represents a different SNR condition. As the SNR increased, the correlation between the envelope of the speech-plus-noise signal and the clean (speech-in-quiet) signal in each spectral band became greater, ultimately reaching a correlation coefficient of 1.0 (no noise or reverberation). These correlations were reflected in the output of the STMI model: with increasing SNR, as the spectral and temporal modulations of the speech-plus-noise envelopes began to resemble the modulations in the "clean" speech envelope, the model predicted an increase in speech intelligibility (Fig. 3.10). To model AV interaction, the visual enhancement was carried out based on dynamic measurements of the two-dimensional positions of 14 reference points on the talker's face made using an OPTOTRAK camera (Fig. 3.8, video data). The 28 resulting visual waveforms (x- and y-coordinates for each transmitter), along with the speech-in-noise envelopes from each frequency channel (cochlear filter), were input as predictor variables into a linear-regression model to predict the clean-speech envelope in each of 136 peripheral frequency bands.

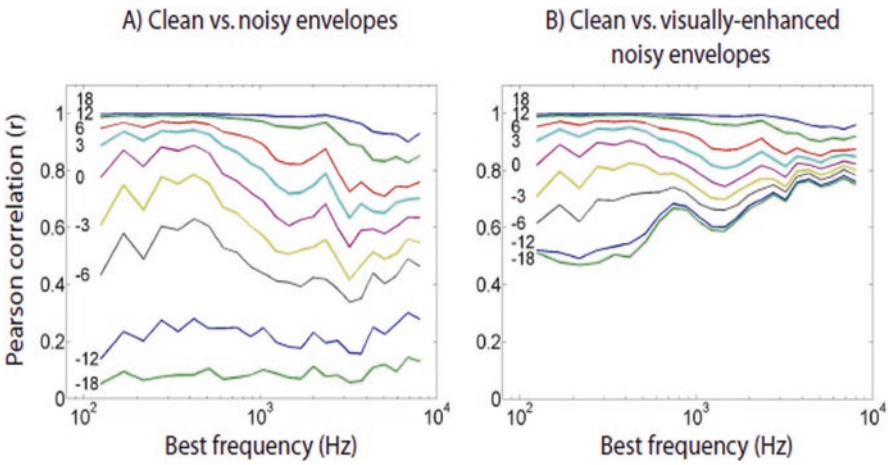

Fig. 3.9 (**A**) Correlation between clean and noisy acoustic speech envelopes for 136 peripheral auditory channels with center frequencies between 125 and 8000 Hz. The speech materials consisted of spoken IEEE sentences. The parameter is the SNR for the A-only speech signal. (**B**) Same as (**A**) except that the speech envelopes were enhanced using visual speech kinematics derived from 14 optical sensors positioned around the lips, cheeks, and chin of the speaker. From Grant et al. (2013)

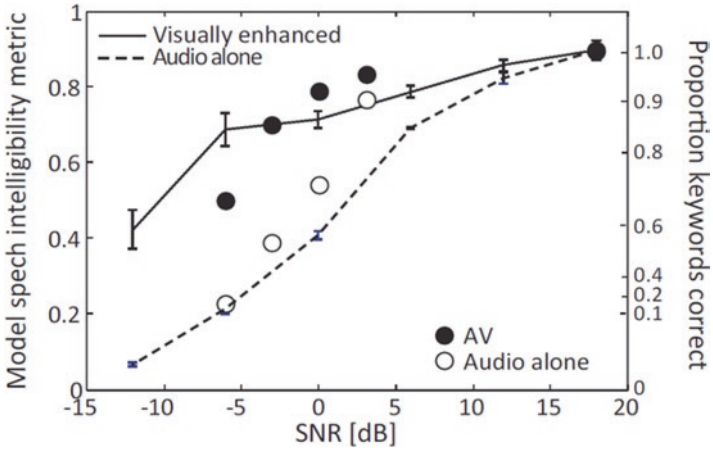

Fig. 3.10 Predicted AV (*solid line*) and A-only (*dashed line*) intelligibility based on the visually enhanced STMI model. *Circles*, intelligibility data measured in normal-hearing listeners Error bars are ±1 SD from model estimates for a list of IEEE sentences processed at each SNR tested

Figure 3.9B shows the correlations between the enhanced speech envelopes (based on the speech-in-noise envelope for each channel and the visual inputs) and the clean-speech envelopes. As in Fig. 3.9A, the correlations generally increased with increasing SNR because the speech-in-noise envelopes became more like the clean-speech envelopes. However, for the AV model, the correlation with the SNR

was considerably higher, especially for low-SNR conditions, than in the A-only case (Fig. 3.9A). This is because the visual speech motion information also contributed to the prediction of the speech envelope. In fact, the AV correlations never decreased below the A-only values obtained for an SNR of approximately −6 dB. At very low SNRs (i.e., −12 and −18 dB), the speech-in-noise acoustic envelopes contained virtually no target speech information, and the prediction was based purely on the visual inputs. Thus, the predicted speech intelligibility was never poorer than that based on the visual channels alone.

By comparing the two panels in Fig. 3.9, it can be seen that the model accounted for the frequency dependence of the AV enhancement similar to what has been observed in perceptual studies (e.g., Grant and Walden 1996). At low frequencies, there was a relatively small difference between the correlations for A-only speech (Fig. 3.9A) and the correlations for AV speech (Fig. 3.9B), meaning that the model showed relatively little visual enhancement to the auditory envelopes when the low-frequency auditory information was corrupted. This is because the visual signal was relatively uninformative (complementary information) about acoustic speech information in this frequency region. In contrast, at high frequencies where the visual signal was predictive of the auditory envelope (redundant information), the visual signal more dramatically enhanced the resulting correlation, meaning that the model showed a large enhancement when high-frequency auditory information was corrupted.

Once the noisy-speech envelopes were enhanced using the temporal dynamics of the visual speech signal to more closely resemble the clean auditory speech envelopes, the cortical and evaluation stages of the model were carried out just as if the envelopes had been generated in the purely acoustic domain but now predicted a higher level of speech intelligibility because the peripheral envelopes more closely resembled clean speech. Figure 3.10 plots the model-predicted speech-intelligibility scores (solid and dashed curves) against the speech-intelligibility scores for sentence materials presented to NH adults (closed and open circles) in speech-shaped noise. The model captured the increase in intelligibility provided by the visual signal as well as the diminishing visual benefit associated with higher SNRs.

The key advantage of this signal-based approach to modeling AV speech intelligibility is that it could successfully account for important aspects of AV speech perception (cf. Sect. 3.1) that traditional models cannot achieve. Although Fig. 3.10 shows that this model captured some of the key features of the relationship between AV benefit and SNR, this is not the same as demonstrating that the model represents an improvement in the ability to predict AV speech intelligibility. In fact, the AI and SII models also predict a contribution of the visual component decreasing with SNR (Fig. 3.2). What this model accomplished beyond the traditional models is (1) the ability to predict AV speech intelligibility based on physical measurements of the speech and noise signal (like the AI, SII, and STI) without requiring a feature-based analysis of auditory- and visual-cue redundancy or an information analysis of A-only and V-only consonant confusions; and (2) an ability to account for spectral interactions when predicting AV speech perception (Fig. 3.9). The model also has the potential to account for AV synchrony effects, although that was not investigated

here. Specifically, the imperviousness of the AV benefit to temporal misalignment (Fig. 3.5) could be modeled by computing a cross-correlation and choosing the delay in each channel that produces the maximum cross-correlation, while adhering to the characteristics of the AV temporal integration window.

3.3 Future Challenges

3.3.1 Complex Auditory Backgrounds

All the AV speech-intelligibility phenomena and modeling (cf. Sects. 3.1 and 3.2) deal with the simple case of NH listeners presented with speech in stationary background noise or filtered speech. In everyday environments, listening situations are much more complex, involving, for example, speech maskers, modulated noise, and spatial separation between target and masker. Although standard speech-intelligibility models (e.g., AI, SII, STI) do not generally address these complex factors, even in A-only situations, substantial research has taken place to understand how these factors influence speech perception in everyday environments. As a result, steps have been taken to incorporate some of these effects into models of auditory speech perception. For example, Rhebergen and Versfeld (2005) and Rhebergen et al. (2006) modified the SII to allow for predictions of speech intelligibility in modulated-noise backgrounds.

Despite the advances made in understanding the complex factors that influence A-only speech perception, relatively little is known about how visual cues interact with spatial cues, variability in masker type, or hearing loss. There have been a handful of studies investigating some of these interactions. For example, Helfer and Freyman (2005) have shown that visual cues can provide an important grouping cue for auditory-scene analysis in multitalker settings, with AV coherence providing the listener with information to perceptually segregate the speech produced by the target talker of interest from a concurrent interfering talker. Bernstein and Grant (2009) found little interaction between hearing loss and the influence of visual cues for speech perception in complex backgrounds. Together, these results suggest that the effects of hearing loss and visual benefit can be modeled independently, but the interaction between the availability of visual information and the perceptual separation of concurrent talkers is likely more complex.

3.3.2 Individual Differences: Hearing Acuity, Visual Acuity, and Integration Efficiency

Several attempts have been made to model the effects of hearing impairment on speech intelligibility (e.g., Bernstein et al. 2013; Bruce 2017). In most of these attempts, only the effects of reduced audibility have been modeled. Individual

differences in spectral and temporal resolution, central auditory processing, and cognitive processing (e.g., working memory, speed of processing, attention allocation), each known to be important for speech intelligibility and understanding, remain a significant challenge (but see Bernstein et al. 2013).

Another area of AV speech perception that would need to be incorporated into any comprehensive model involves degradation in the visual domain due to vision loss (Hardick et al. 1970). Although typical age-related vision loss does not eliminate the visual speech-intelligibility benefit (Hickson et al. 2004), blurred vision can reduce the effect (Legault et al. 2010). Evidence from earlier studies suggests that speechreading performance significantly declines with age, especially for those over 70 years old (Shoop and Binnie 1979; Middelweerd and Plomp 1987). Although the reasons for this decline are not fully understood, it has been suggested that reductions in peripheral visual acuity and motion perception associated with aging may play a role. Unfortunately, there are very few studies that have examined the relationship between overall speechreading ability, individual differences in the transmission of visual place-of-articulation information, and visual acuity. Therefore, if the goal is to predict AV speech intelligibility as well as individual differences in AV processing due to hearing and vision loss, basic tests of auditory and visual function will have to be incorporated into the modeling efforts.

Finally, there is the possibility that some individuals are better able than others to integrate auditory and visual information. As discussed in Sect. 3.1.4, although many of the differences in AV benefit observed by HI listeners can be ascribed to an impoverished auditory signal, there was at least some evidence that certain individuals might also have had a particular deficit in the ability to integrate speech information from the two modalities (Grant et al. 2007). To the extent that individual variability in integration efficiency exists, this factor would also need to be included in an individual-specific model of AV speech perception.

3.4 Summary

Signal-based models of speech perception are critically important for the design and evaluation of audio systems and hearing-rehabilitation devices. Models such as the AI, SII, and STI have undergone decades of development and scrutiny and are mostly successful in predicting average speech intelligibility for acoustic signals under a variety of conditions. Yet more than half of speech communication takes place in face-to-face situation where the listener is looking at the talker and has access to visual speech cues (Walden et al. 2004). It is clear that the simplistic approach in the manner in which that these models predict AV speech intelligibility, assuming that the speechreading benefit to auditory speech intelligibility can be modeled as a simple additive factor, is incorrect. Thus, these extant models are inadequate for predicting AV speech intelligibility for a given audio input signal, transducer, and hearing loss.

Section 3.1 reviewed several important phenomena associated with AV speech perception that highlight the complex interaction between these modalities that any model would need to take into account. In particular, it was shown that the amount of speechreading benefit depends dramatically on the spectral content of the speech signal (Grant et al. 1985; Grant and Walden 1996). This interaction can be understood in terms of the complementary or redundant nature of the speech features provided by the visual and acoustic speech cues (Grant and Walden 1996). Although extant models of speech-feature integration proposed by Massaro (1987) and Braida (1991) do a good job predicting AV speech recognition for nonsense syllables, they cannot predict sentence or connected discourse performance and require significant time and effort to obtain unimodal perceptual confusion-matrix data. Other important aspects of AV speech perception that the simple additive models cannot account for include a limited tolerance to temporal asynchrony within a range of −20 ms (audio leading) to +200 ms (audio lagging) (Grant et al. 2004; Shahin et al. 2017) and the possibility of individual variability in AV integration efficiency (Grant et al. 2007).

Section 3.2 described a signal-based modeling approach to predicting AV speech perception. One of the greatest obstacles to developing a model of AV speech perception has been the centuries-old tradition of treating sensory modalities as independent receivers of information, combined at an abstract linguistic level. However, physiological data showing the existence of multimodal neurons that only fire when certain temporal constraints across inputs from different sensory modalities are met suggest a different story. In fact, listeners are sensitive to coherence in the modulation of the acoustic envelope and the temporal dynamics of lip movements (Grant and Seitz 2000; Grant 2001), providing a clue for how AV speech performance might be predicted from the physical properties of the visual and acoustic signals. In the model, visual speech motion was used to help reconstruct and enhance corrupted auditory speech-envelope information from different frequency channels into envelopes that more closely resemble those from clean speech. This approach was shown to be consistent with experimental evidence that the visual signal is best able to stand in for corrupted acoustic speech information in the mid-to-high speech frequencies associated with F2 and F3 transitions (place-of-articulation information). Although work remains to integrate other important known aspects of AV speech processing (e.g., tolerance to asynchrony, individual variation in visual or hearing acuity, and integration efficiency), this approach represents an important step toward the development of a signal-based AV speech-perception model in the spirit of the AI, SII, and STI.

Compliance with Ethics Requirements Kenneth W. Grant declares that he has no conflict of interest.

Joshua G. W. Bernstein declares that he has no conflict of interest.

The views expressed in this chapter are those of the authors and do not reflect the official policy of the Department of Army/Navy/Air Force, Department of Defense, or US Government.

References

American National Standards Institute (ANSI). (1969). *American National Standard Methods for the calculation of the articulation index. ANSI S3.5-1969*. New York: American National Standards Institute.

American National Standards Institute (ANSI). (1997). *American National Standard Methods for calculation of the speech intelligibility index. ANSI S3.5–1997*. New York: American National Standards Institute.

Bernstein, J. G. W., & Grant, K. W. (2009). Audio and audiovisual speech intelligibility in fluctuating maskers by normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America, 125*, 3358–3372.

Bernstein, J. G. W., Summers, V., Grassi, E., & Grant, K. W. (2013). Auditory models of suprathreshold distortion and speech intelligibility in persons with impaired hearing. *Journal of the American Academy of Audiology, 24*, 307–328.

Berthommier, F. (2004). A phonetically neutral model of the low-level audio-visual interaction. *Speech Communication, 44*(1), 31–41.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology, 43*, 647–677.

Bruce, I. (2017). Physiologically based predictors of speech intelligibility. *Acoustics Today, 13*(1), 28–35.

Byrne, D., Dillon, H., Ching, T., Katsch, R., & Keidser, G. (2001). NAL-NL1 procedure for fitting nonlinear hearing aids: Characteristics and comparisons with other procedures. *Journal of the American Academy of Audiology, 31*, 37–51.

Drullman, R., & Smoorenburg, G. F. (1997). Audio-visual perception of compressed speech by profoundly hearing-impaired subjects. *Audiology, 36*(3), 165–177.

Elhilali, M., Chi, T., & Shamma, S. A. (2003). A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Communication, 41*(2), 331–348.

Erber, N. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech, Language, and Hearing Research, 15*(2), 413–422.

Fletcher, H. (1953). *Speech and hearing in communication*. New York: Van Nostrand.

Fletcher, H., & Gault, R. H. (1950). The perception of speech and its relation to telephony. *The Journal of the Acoustical Society of America, 22*, 89–150.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America, 19*, 90–119.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., et al. (1990). *DARPA, TIMIT acoustic-phonetic continuous speech corpus CD-ROM*. Gaithersburg, MD: National Institute of Standards and Technology, US Department of Commerce.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1*. Gaithersburg, MD: National Institute of Standards and Technology, US Department of Commerce.

Girin, L., Schwartz, J. L., & Feng, G. (2001). Audio-visual enhancement of speech in noise. *The Journal of the Acoustical Society of America, 109*(6), 3007–3020.

Gordon, P. C. (1997). Coherence masking protection in speech sounds: The role of formant synchrony. *Perception & Psychophysics, 59*, 232–242.

Gordon, P. C. (2000). Masking protection in the perception of auditory objects. *Speech Communication, 30*, 197–206.

Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America, 109*, 2272–2275.

Grant, K. W., Ardell, L. H., Kuhl, P. K., & Sparks, D. W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *The Journal of the Acoustical Society of America, 77*, 671–677.

Grant, K. W., Bernstein, J. G. W., & Grassi, E. (2008). Modeling auditory and auditory-visual speech intelligibility: Challenges and possible solutions. *Proceedings of the International Symposium on Auditory and Audiological Research, 1*, 47–58.

Grant, K. W., Bernstein, J. G. W., & Summers, V. (2013). Predicting speech intelligibility by individual hearing-impaired listeners: The path forward. *Journal of the American Academy of Audiology, 24*, 329–336.

Grant, K. W., & Braida, L. D. (1991). Evaluating the articulation index for audiovisual input. *The Journal of the Acoustical Society of America, 89*, 2952–2960.

Grant, K. W., Greenberg, S., Poeppel, D., & van Wassenhove, V. (2004). Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing. *Seminars in Hearing, 25*, 241–255.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*, 1197–1208.

Grant, K. W., Tufts, J. B., & Greenberg, S. (2007). Integration efficiency for speech perception within and across sensory modalities. *The Journal of the Acoustical Society of America, 121*, 1164–1176.

Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for auditory-visual consonant recognition. *The Journal of the Acoustical Society of America, 100*, 2415–2424.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America, 103*, 2677–2690.

Hall, J. W., Haggard, M. P., & Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. *The Journal of the Acoustical Society of America, 76*, 50–56.

Hardick, E. J., Oyer, H. J., & Irion, P. E. (1970). Lipreading performance as related to measurements of vision. *Journal of Speech and Hearing Research, 13*, 92–100.

Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *The Journal of the Acoustical Society of America, 117*(2), 842–849.

Hickson, L., Hollins, M., Lind, C., Worrall, L. E., & Lovie-Kitchin, J. (2004). Auditory-visual speech perception in older people: The effect of visual acuity. *Australian and New Zealand Journal of Audiology, 26*, 3–11.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *The Journal of the Acoustical Society of America, 73*(1), 322–335.

Killion, M., Schulein, R., Christensen, L., Fabry, D., Revit, L., Niquette, P., & Chung, K. (1998). Real-world performance of an ITE directional microphone. *The Hearing Journal, 51*, 24–39.

Legault, I., Gagné, J. P., Rhoualem, W., & Anderson-Gosselin, P. (2010). The effects of blurred vision on auditory-visual speech perception in younger and older adults. *International Journal of Audiology, 49*(12), 904–911.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431–461.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *The Journal of the Acoustical Society of America, 100*(3), 1777–1786.

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *The Journal of the Acoustical Society of America, 77*(2), 678–685.

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America, 82*(6), 2145–2147.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America, 27*(2), 338–352.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time. *Speech Communication, 41*(1), 245–255.

Reetz, H., & Jongman, A. (2011). *Phonetics: Transcription, production, acoustics, and perception*. Chichester, West Sussex: Wiley-Blackwell.

Rhebergen, K. S., & Versfeld, N. J. (2005). A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America, 117*(4), 2181–2192.

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *The Journal of the Acoustical Society of America, 120*(6), 3988–3997.

Rosen, S. M., Fourcin, A. J., & Moore, B. C. J. (1981). Voice pitch as an aid to lipreading. *Nature, 291*(5811), 150–152.

Shahin, A. J., Shen, S., & Kerlin, J. R. (2017). Tolerance for audiovisual asynchrony is enhanced by the spectrotemporal fidelity of the speaker's mouth movements and speech. *Language, Cognition and Neuroscience, 32*(9), 1102–1118.

Shoop, C., & Binnie, C. A. (1979). The effects of age upon the visual perception of speech. *Scandinavian Audiology, 8*(1), 3–8.

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26*(3), 263–275.

Steeneken, H. J., & Houtgast, T. (2002). Validation of the revised STI_r method. *Speech Communication, 38*(3), 413–425.

Studdert-Kennedy, M. (1974). The perception of speech. *Current Trends in Linguistics, 12*, 2349–2385.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*, 212–215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–52). Hillsdale NJ: Lawrence Erlbaum Associates.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London B, Biological Sciences, 335*(1273), 71–78.

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing, 28*(5), 656–668.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 1181–1186.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*, 598–607.

Walden, B. E., Grant, K. W., & Cord, M. T. (2001). Effects of amplification and speechreading on consonant recognition by persons with impaired hearing. *Ear and Hearing, 22*(4), 333–341.

Walden, B. E., Surr, R. K., Cord, M. T., & Dyrlund, O. (2004). Predicting hearing aid microphone preference in everyday listening. *Journal of the American Academy of Audiology, 15*(5), 365–396.

Wu, Y. H., & Bentler, R. A. (2010). Impact of visual cues on directional benefit and preference: Part I—Laboratory tests. *Ear and Hearing, 31*(1), 22–34.