Language Teaching (2021), **54**, 206–220
doi:10.1017/S0261444820000592

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH TIMELINE

Multimodal input in second-language speech processing

Debra M. Hardison*

Michigan State University, East Lansing, USA
*Corresponding author. Email: hardiso2@msu.edu

Introduction

This timeline provides an update on research since 2009 involving auditory-visual (AV) input in spoken language processing (see Hardison, 2010 for an earlier timeline on this topic). A brief background is presented here as a foundation for the more recent studies of speech as a multimodal phenomenon (e.g., Rosenblum, 2005).

In the 1950s, some researchers suggested that the prevailing view of speech as an auditory-only (A-only) event had overlooked an important source of input. Sumby and Pollack (1954) argued that speech intelligibility could be enhanced by observation of the speaker; specifically, lip movements could be helpful for disambiguating consonant sounds (Miller & Nicely, 1955). Subsequent studies demonstrated that visual cues from a speaker's face offered an advantage in the accurate identification of speech sounds for a variety of listener populations, languages, and stimulus conditions. These included the hearing impaired (e.g., Walden, Prosek, Montgomery, Scherr, & Jones, 1977; Bergeson, Pisoni, & Davis, 2003), non-impaired listeners trying to comprehend conceptually difficult messages or accented speech (Reisberg, McLean, & Goldfield, 1987), non-impaired listeners in ambient noise (e.g., Benoît, Mohamadi, & Kandel, 1994 for French; MacLeod & Summerfield, 1990 for English), listeners of speech presented in the clear (e.g., McGrath & Summerfield, 1985), infants in their first language (L1) development (e.g., Meltzoff & Moore, 1993), and individuals responding to the McGurk Effect (e.g., McGurk & MacDonald, 1976; Hardison, 1996) – a perceptual effect in which discrepant AV cues may result in an illusory percept (e.g., a combination of auditory /ba/ and visual /ga/ may produce the percept /da/). For second-language (L2) learners of English (L1 Japanese and Korean), training with visual cues from a native speaker's face improved their identification accuracy of /r/ and /l/ which transferred to production improvement (Hardison, 2003) and earlier identification of words beginning with those sounds (Hardison, 2005). Articulatory gestures often precede the associated acoustic signal, essentially giving the listener/observer a head start in reducing the set of potential candidates in the speech recognition process (e.g., Munhall & Tohkura, 1998). In addition to the information value of lip movements in face-to-face interactions, head and eyebrow movements, which are correlated with changes in vocal pitch, also improve speech perception (e.g., Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). Among other nonverbal cues, hand-arm gestures contribute to language comprehension (Sueyoshi & Hardison, 2005; Gullberg, 2006); in particular, beat gestures can focus perceivers' attention on certain elements in multimodal discourse (Dimitrova, Chu, Wang, Özyürek, & Hagoort, 2016).

Early demonstrations of the advantage of computer-based visual displays of acoustic information in L2 training involved L1 Dutch speakers learning Chinese lexical tones (Leather, 1990) and English sentence-level intonation (de Bot, 1983). Further advances in computer-based sources of visual feedback on one's own speech or that of a model speaker led to widespread use of the acronym CAPT (Computer-Assisted Pronunciation Training; see Hincks, 2015 for a review). This feedback includes displays of waveforms, which can visually represent the duration of sounds (e.g., Motohashi-Saigo & Hardison, 2009); spectrograms, which show the internal structure of a sound's acoustic energy (e.g., Hardison, 2019); and pitch tracking for visualizing the rise and fall of vocal pitch (e.g.,

© The Author(s), 2020. Published by Cambridge University Press

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

Hardison, 2004). Other displays of articulatory information, such as ultrasound imaging with its history in clinical applications, have been adapted to provide feedback in language learning (e.g., Gick, Bernhardt, Bacsfalvi, & Wilson, 2008). Automatic Speech Recognition (ASR) systems that rely on machine learning can be used for feedback through dictation programs or segmental error detection in a learner's utterance (e.g., O'Brien et al., 2018). Computer-animated faces or talking heads (e.g., Massaro & Light, 2003) have the potential to serve as virtual pronunciation teachers with the goal of showing the articulatory movements necessary to produce a target sound.

The importance and contribution of multimodal input is supported by the findings of neuroimaging studies which have revealed a relationship between visual input and activity in the auditory cortex during speech perception (e.g., Sams et al., 1991; Hickok, Buchsbau, Humphries, & Muftuler, 2003). The developmental role of multimodal input is captured well in the following statement from Shams and Seitz (2008), 'It is likely that the human brain has evolved to develop, learn and operate optimally in multisensory environments.' (p. 411).

The updated timeline on spoken language processing is constructed around the following themes based on the studies that have informed the directions that multimodal speech research has been taking.

1. Input from the human face
2. Electronic displays
    - **A.** Displays of acoustic information (e.g., pitch, waveforms)
    - **B.** Displays of articulatory position and movement (e.g., ultrasound)
    - **C.** Computer-animations or talking heads
    - **D.** ASR
3. Hand-arm gestures
4. Neurophysiological studies of AV processing

References

Benoît, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research*, *37*(5), 1195–1203.

Bergeson, T. R., Pisoni, D. B., & Davis, R. A. O. (2003). A longitudinal study of audiovisual speech perception by children with hearing loss who have cochlear implants. *Volta Review*, *103*(4), 347–370.

de Bot, K. (1983). Visual feedback of intonation I: Effectiveness and induced practice behaviour. *Language and Speech*, *26*(4), 331–350.

Dimitrova, D., Chu, M., Wang, L., Özyürek, A., & Hagoort, P. (2016). Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience*, *28*(9), 1255–1269.

Gick, B., Bernhardt, B. M., Bacsfalvi, P., & Wilson, I. (2008). Ultrasound imaging applications in second language acquisition. In J. Hansen & M. Zampini (Eds.), *Phonology and Second Language Acquisition* (pp. 309–322). Amsterdam, Netherlands: John Benjamins.

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *IRAL*, *44*(2), 103–124.

Hardison, D. M. (1996). Bimodal speech perception by native and non-native speakers of English: Factors influencing the McGurk effect. *Language Learning*, *49*(1), 213–283.

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*(4), 495–522.

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, *8*(1), 34–52.

Hardison, D. M. (2005). Second-language spoken word identification: Effects of perception training, visual cues, and phonetic environment. *Applied Psycholinguistics*, *26*(4), 579–596.

Hardison, D. M. (2010). Visual and auditory input in second-language speech processing. *Language Teaching*, *43*(1), 84–95.

Hardison, D. M. (2019). Technology-based communication success for second-language learners. In M. A. Peters & R. Heraud (Eds.), *Encyclopedia of Educational Innovation*. Singapore: Springer: doi.org/10.1007/978-981-13-2262-4_85-2

Hickok, G. B., Buchsbau, B., Humphries, C., & Muftuler, T. (2003). Auditory-motor interaction revealed by fMRI: Speech, music, and working memory in area Spt. *Journal of Cognitive Neuroscience*, *15*(5), 673–682.

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

Hincks, R. (2015). Technology and learning pronunciation. In M. Reed & J. M. Levis (Eds.), *The handbook of English pronunciation* (pp. 505–519). Hoboken, NJ: Wiley.
Leather, J. (1990). Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers. In J. Leather & A. James (Eds.), *New Sounds 90* (pp. 72–97). Amsterdam, Netherlands: University of Amsterdam.
MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology*, *24*(1), 29–43.
Massaro, D. W., & Light, J. (2003). Read my tongue movements: Bimodal learning to perceive and produce non-native speech /r/ and /l/. In *EUROSPEECH-2003*, pp. 2249–2252.
McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, *77*(2), 678–685.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.
Meltzoff, A., & Moore, M. K. (1993). Why faces are special to infants – on connecting the attraction of faces and infants' ability for imitation and cross-modal processing. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. McNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 211–225). Dordrecht, Netherlands: Kluwer Academic.
Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, *27*, 338–352.
Motohashi-Saigo, M., & Hardison, D. M. (2009). Acquisition of L2 Japanese geminates: Training with waveform displays. *Language Learning & Technology*, *13*(2), 29–47.
Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133–137.
Munhall, K. G., & Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America*, *104*, 530–539.
O'Brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. L., … Levis, G. M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation*, *4*(2), 182–206.
Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). London, UK: Erlbaum.
Rosenblum, L. (2005). The primacy of multimodal speech perception. In D. Pisoni & R. Remez (Eds.), *Handbook of Speech Perception* (pp. 51–78). Malden, MA: Blackwell.
Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S.-T., & Simola, J. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*(1), 141–145.
Shams, L. A., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–417.
Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second-language listening comprehension. *Language Learning*, *55*(4), 661–699.
Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215.
Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, *20*(1), 130–145.

Debra M. Hardison is affiliated with the Department of Linguistics and Languages at Michigan State University. She conducts research on multimodal integration in spoken language processing, co-speech gesture, factors affecting oral skills development for L2 learners, and the applications of technology in L2 perception and production training involving segmental and suprasegmental features. She has given plenaries at *New Sounds: International Symposium on the Acquisition of Second Language Speech* and *Pronunciation in Second Language Learning & Teaching* and publishes in numerous journals and edited collections. She teaches graduate courses in L2 phonetics, and research methods in language learning and teaching.

Cite this article: Hardison, D. M. (2021). Multimodal input in second-language speech processing. *Language Teaching*, *54*(2), 206–220. https://doi.org/10.1017/S0261444820000592

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

Year	References	Annotations	Theme
2009	Cucchiarini, C., Neri, A. & Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication*, *51*(10), 853–863.	To explore the pedagogical effectiveness of automatic speech recognition technology, *Dutch-CAPT* was tested as a means of providing corrective feedback on 11 L2 Dutch segmental pronunciation errors made by beginner-level learners. Although learners found the training (2 sessions, each 30–60 minutes) useful and became aware of pronunciation problems, the system's accuracy in error detection was 75–86%, and it did not provide feedback on a large number of problem areas. Even though learners' pronunciation did not improve significantly, the number of mispronunciations was much smaller for the experimental (vs. control) group.	**B3**
2009	Kawase, T., Sakamoto, S., Hori, Y., Maki, A., Suzuki, Y., & Kobayashi, T. (2009). Bimodal audio-visual training enhances auditory adaptation process. *NeuroReport*, *20*, 1231–1234.	**Kawase et al.** trained L1 Japanese listeners with either A-only or AV presentation of Japanese words in which highly degraded noise-vocoded speech was used as a simulation of cochlear implant speech. Results showed that visual cues improved perception of both trained and untrained (novel) degraded speech sounds, suggesting that visual cues enhanced the auditory adaptation process, which bodes well for populations such as the hearing impaired and language learners who experience adaptation needs.	**A**
2009	Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics*, *37*, 344–356.	Following L2 perception studies demonstrating the benefits of AV speech for non-native listeners (e.g., Hardison, 1996*), **Wang et al.** used A-only, V-only (video-only), and congruent and incongruent AV presentation of English CV (consonant-vowel) syllables to L1 Korean, Mandarin, and English speakers. Stimuli contained fricatives at different points of articulation: labiodentals (nonexistent in Korean), interdentals (nonexistent in Korean and Mandarin), and alveolars (present in all L1s). Results indicated lower V-only accuracy of labiodentals for Korean speakers, and lower A-only accuracy of interdentals for Korean and Mandarin speakers. Findings provided further support for the influence of L1 background on the degree to which individuals can use L2 auditory and visual information.	**A**
2010	de Barros, P. C. M. (2010). 'It's easier to understand': The effect of a speaker's accent, visual cues, and background knowledge on listening comprehension (Master's thesis). Kansas State University.	This study further explored the facilitating effect of a lecturer's facial cues and gestures on L2 English comprehension (Sueyoshi & Hardison, 2005*). An introductory level entomology lecture was given by an English native speaker and a non-native speaker (L1 Portuguese teaching assistant) in AV and A-only conditions to L1 English college-level students. Comprehension scores revealed that visual cues (gestures and facial expressions) provided marginal facilitation although participants tended to exhibit more difficulty comprehending the non-native speaker in the A-only condition despite having knowledge of her L1. It was noted that visual cues were not required for comprehension of the topic. Topic familiarity played a role, especially in understanding the non-native lecturer. Neither a description of the lecturers' gestures nor the comprehensibility of the L2 English was provided.	**A, C**

(*Continued*)

*denotes the full reference appears in the references list above.

Note. Authors' names are shown in small capitals when the study referred to appears in this timeline.

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

(*Continued*)

Year	References	Annotations	Theme
2010	Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication*, *52*, 525–532.	Studies had shown that identification of speech sounds presented in CV syllables could be enhanced by visual cues (e.g., for English, McGurk & MacDonald, 1976*; Hardison, 1996*, 2003*; for French, Benoît et al., 1994*). **Fort et al.** used a phoneme monitoring task with real French words and pseudowords in A-only and AV (speaker's face) conditions with two levels of background noise. Native French speakers detected the target consonants faster in the AV versus A-only condition. With masking noise, consonants were better detected when they were embedded in real (vs. pseudo-) words in the AV condition, suggesting that phoneme detection can be influenced by lexical knowledge and visual information contributes to lexical activation processes.	**A**
2010	Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication*, *52*, 996–1009.	**Hazan et al.** presented CV syllables /ba/, /da/, and /ga/ produced by Australian English and Mandarin speakers to Australian English, British English, and Mandarin listeners in several conditions: A-only, V-only, congruent and incongruent AV, either in the clear, in noise, with blurring or combined AV degradations. Results indicated that while non-native speaker and L1 background influenced the weighting of auditory and visual cues as suggested by Wang et al. (2009), cue weighting varied across participants and was influenced by individual speaker characteristics.	**A**
2010	Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*, *53*(2), 298–310.	**Hirata and Kelly** explored whether a Japanese speaker's hand and/or mouth movements would aid perception of Japanese vowel length contrasts by L1 English speakers with no prior exposure to Japanese. Using a pretest (A-only)-training (four sessions)-posttest (A-only) design, four training conditions were created: audio, audio + mouth, audio + hand (face obscured but hand visible), audio + mouth + hand (all movements visible). Participants were told that short vertical and long horizontal hand movements corresponded to short and long vowels respectively. All training groups improved, but only the audio + mouth condition was significantly better than audio-only in contrast to Sueyoshi and Hardison (2005*), who found perceptual benefit for L2 English learners presented with articulatory and hand gestures (see also Hardison, 2019).	**A, C**
2010	Jesse, A., & Massaro, D. W. (2010). Seeing a singer helps comprehension of the song's lyrics. *Psychonomic Bulletin & Review*, *17*(3), 323–328.	This study demonstrated that understanding the lyrics of a song benefits from seeing the singer, similar to findings from numerous studies that have shown that seeing a speaker's face facilitates comprehension of speech for native and non-native listeners (e.g., Hardison, 2003*; Hardison, 2018a; Hirata & Kelly, 2010). Findings indicated a word recognition improvement of about 35% for AV presentation, which is striking given that singing often alters articulatory gestures, and thus could be more obstructive to comprehension than facilitative. However, sung lyrics offer several potential benefits: place-of-articulation information for consonants, lengthened vowel duration, and the correlation between head and eyebrow movements with pitch changes (e.g., Hardison, 2018b; Munhall et al., 2004, see Introduction). Findings support the	**A**

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

		notion that both speech and music perception are multisensory processes (e.g., Rosenblum, 2005*)	
2010	Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., & Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. *Brain Research*, *1323*, 84–93.	Studies have shown that articulatory gestures precede the associated acoustic signal (Munhall & Tohkura, 1998*), and that such a temporal precedence can serve a priming role for a listener/observer in the speech recognition process. This role depends on the salience of the visual cue and the perceiver's language experience (e.g., Hardison, 1996*, 2005*; Hardison, 2018a). **Navarra et al.** presented AV clips recorded in English and Spanish by a bilingual speaker to native speakers of both of those languages who had little prior experience with the non-native language. The cues were presented at varying onset asynchronies, and participants were asked to determine if the AV cues were synchronous or not. Results indicated that to achieve perceived synchrony, a larger lead for the visual cue was required in the L1 than in the unfamiliar language, regardless of the language or participant group. In contrast, Spanish speakers who were more proficient in English showed no difference between the native and non-native language.	**A**
2011	Ruellot, V. (2011). Computer-assisted pronunciation learning of French /u/ and /y/ at the intermediate level. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 2nd Pronunciation in Second Language Learning and Teaching Conference* (pp. 199–213). Ames, IA: Iowa State University.	In this study comparing AV and A-only input, side-by-side spectrograms of French [y] and [u] were used as visual feedback for a group of L1 English learners of French in a phonetics course to help them differentiate these two vowels in their pronunciation. The primary formants of each vowel were color-coded and learners' production task was to approximate the distances between the formants in the native speakers' examples. Both the AV and A-only groups showed significant production improvement; however, no additional benefit was found for those who had seen the spectrograms (see Offerman & Olson, 2016; Olson, 2014).	**B1**
2011	Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *Calico Journal*, *28*(3), 744–765.	Mandarin speakers were trained to associate a distinctive nautical flag image with each of ten target English vowels. In response to a native-speaker's recording of each vowel in a syllable, learners selected the appropriate image on a computer screen to indicate the vowel they had heard. Visual and auditory feedback was given. Training resulted in significant improvement in the intelligibility of their vowel productions, consistent with other studies demonstrating a perception-production link (e.g., Hardison, 2003*; Okuno and Hardison, 2016).	**B**
2012	Engwall, O. (2012). Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. *Computer Assisted Language Learning*, *25*(1), 37–64.	Talking heads such as *Baldi* (e.g., Massaro & Light, 2003*) have been used in speech production training in clinical applications (e.g., with the hearing impaired) and L2 learning. Expanding on this technology, **Engwall** used ultrasound imaging to monitor learners' articulation change following the AV feedback instructions given in a short session by *Artur*, a semi-transparent 3D tutor that shows articulatory movements. This feedback includes the relationship between the tongue and palate in sound production. French speakers practiced the pronunciation of the Swedish alveolar trill [r] and a velar fricative. Animations from *Artur* demonstrated the articulation change that was given in the instructions. Results showed some evidence that participants could change their articulations according to the instructions on how to reposition the tongue.	**B2, 3**

(*Continued*)

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

(*Continued*)

Year	References	Annotations	Theme
2012	Schmidt, A. M. (2012). Effects of EPG treatment for English consonant contrasts on L2 perception and production. *Clinical Linguistics & Phonetics*, *26*(11–12), 909–925.	**Schmidt** focused on training two L1 Korean speakers to produce a consistent phonological and articulatory contrast between the members of the following L2 English pairs using electropalatography as visual feedback: [s] – [ʃ], [z] – [ʤ], [l] – [ɹ]. During training sessions using the sounds in initial position in syllables and words, participants wore pseudopalates which produced a computer screen display in which areas turned green when the desired tongue-palate contact was made, or red for inaccurate contact. Results indicated more stable production after training with some transfer to improved perception ability.	**B2**
2012	Strik, H., Colpaert, J., van Doremalen, J., & Cucchiarini, C. (2012). The *DISCO* ASR-based CALL system: Practicing L2 oral skills and beyond. *Proceedings of the Conference on International Language Resources and Evaluation 2012*, Istanbul, Turkey. http://www.lrecconf.org/proceedings/lrec2012/pdf/787_Paper.pdf	In the *Development and Integration of Speech Technology into Courseware (DISCO)* system, ASR technology was used to provide practice and feedback on pronunciation, morphology, and syntax for L2 Dutch learners. In previous research (Cucchiarini et al., 2009), exercises and learners' possible answers had been limited; however, in order for *DISCO* to address frequent and salient morphological, syntactic and pronunciation errors in low proficiency learners' speech, exercises in *DISCO* had to allow learners more freedom in producing answers, but still remain predictable by ASR modules. Learners chose a prompt from the screen to read in a simulated real-world conversation with an animated agent with lip movements that move in synchrony with recorded utterances. Pronunciation feedback was available through the dialogues and in remedial exercises. Learners chose either explicit feedback through a visually highlighted error, which stops the dialogue, or a recast with emphasis on the correct form through highlighting, allowing the dialogue to continue. At the end, learners received an error summary and the option to practice problematic features.	**B3, 4**
2012	Wang, X. (2012). Auditory and visual training on Mandarin tones: A pilot study on phrases and sentences. *International Journal of Computer-Assisted Language Learning and Teaching*, *2*(2), 16–29.	Earlier research had focused on Mandarin lexical tones (e.g., Leather, 1990*). In this study, beginning-level learners of Mandarin from a range of L1 backgrounds received six hours of training focused on tones in phrases and sentences using real-time computer displays of pitch contours to compare their productions with those of multiple native speakers. Learners' tone production significantly improved following training.	**B1**
2013	Moradi, S., Lidestam, B., & Rönnberg, J. (2013). Gated audiovisual speech identification in silence vs. noise: Effects on time and accuracy. *Frontiers in Psychology*, *4*, 359.	Using the gating paradigm, Hardison (2005*; Hardison, 2018a) found that visual cues (seeing a speaker's face) facilitated word identification by native speakers and L2 learners of English compared to A-only presentation. In gating, successively increasing amounts of a speech stimulus are presented until the target is correctly identified. Using the same paradigm with Swedish consonants in VCV (vowel-consonant-vowel) syllables, **Moradi et al.** found that native speakers of Swedish were faster and more accurate in consonant identification in AV (vs. A-only) presentation in silence and noise; the more visually salient consonants such as /b, f, m, p/ were more resistant to noise. Results corroborate the findings of Fort et al. (2010).	**A**

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

2013	Yi, H-G., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for non-native speech. *Journal of the Acoustical Society of America*, *134*(5), EL387-393.	Previous studies had shown the advantage of visual cues from a speaker's face on speech perception in noise (e.g., MacLeod & Summerfield, 1990*) and perception by native and non-native listeners (e.g., Hardison, 2003*, 2005*). Yi et al. investigated the effect of seeing a speaker's face on L1 English listeners' ability to identify keywords in English sentences produced by native English versus native Korean speakers in noise (background of babble). Keywords were more correctly identified in the AV versus A-only condition, and visual cues benefited perception of speech produced by the native versus the non-native speaker. In a separate accentedness rating task, Korean speakers were rated as more strongly accented in the AV versus A-only condition, suggesting a possible visual bias. (See also Hardison, 2018a; Kawase et al., 2014; Zheng & Samuel, 2019.)	**A**
2014	Hirata, Y., Kelly, S. D., Huang, J., & Manansala, M. (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *Journal of Speech, Language, and Hearing Research*, *57*, 2090–2101.	For native Japanese speakers and listeners/observers, gestures correspond to the moraic rhythm of the language. In this study, L1 English speakers with no knowledge of Japanese were trained on Japanese vowel duration contrasts. In an extension of an earlier study (Hirata & Kelly, 2010), in **Hirata et al.**, participants both observed and produced the gestures. In a pretest (A-only)-training (four sessions)-posttest (A-only) design, training stimuli consisted of pairs of Japanese words contrasting vowel length in the first or second syllable. Native co-speech gesture varied as follows: Words with short vowels were produced with two small downward hand movements ('syllable-rhythm'); those with long vowels were accompanied by either (a) one horizontal dip, followed or preceded by a short vowel with a short movement ('syllable-rhythm') or (b) a total of three movements corresponding to the number of morae in each word ('mora-rhythm'). Along with auditory input, training involved one of the following: observation of the syllable-rhythm gesture, production of syllable-rhythm with the instructor, observation of the mora-rhythm gesture, or production of the mora-rhythm gesture. Results indicated improvement in auditory perception of vowel length for all training types; however, observing the syllable-rhythm gesture typical of L1 English produced the most improvement across word-initial and word-final vowels. (See also Hardison, 2019; Okuno & Hardison, 2016.)	**C**
2014	Katz, W., Campbell, T., Wang, J., Farrar, E., Coleman Eubanks, J., Balasubramanian, A., Prabhakaran, B. & Rennaker, R. (2014). Opti-Speech: A real-time 3D visual feedback system for speech training. In *Proceedings of Interspeech 2014, Singapore*. 1174–1178.	To help speakers reach consonant place-of-articulation targets, this interactive 3D system provided real-time feedback on their tongue and jaw movements based on tongue sensors and a head movement sensor worn on glasses. Preliminary data showed reliable feedback for American English consonant targets. (See also Bliss et al., 2018; Engwall, 2012; Kocjančič Antolík, 2019 for reports using ultrasound as direct articulatory feedback.) These studies have implications for L2 learners and individuals with communication disorders.	**B2**
2014	Kawase, S., Hannah, B., & Wang, Y. (2014). The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers. *Journal of the*	**Kawase et al.** investigated the contribution of visual cues from a speaker's face to native English listeners' perceptions of the intelligibility of CV syllables produced by both native Japanese and native English speakers. Consonant contrasts included /b-v/, /θ-s/, /l-ɹ/ in three stimulus conditions: AV, A-only, and V-only. Findings revealed that intelligibility of the Japanese speakers' productions of /b, s, l/ was better compared to	**A**

(*Continued*)

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

(*Continued*)

Year	References	Annotations	Theme
	Acoustical Society of America, *136*(3), 1352–1362.	/v, θ, ɹ/. Visual cues significantly improved intelligibility for /v, θ/ but not /ɹ/; the non-natives' articulatory gestures lacked the lip protrusion present in native English speakers' productions. The contribution of visual cues to a percept is subject to perceivers' familiarity with the articulatory gestures and their discernibility (Hardison, 1996*) and may be most beneficial with native versus non-native speech (Yi et al., 2013).	
2014	Olson, D. J. (2014). Benefits of visual feedback on segmental production in the L2 classroom. *Language Learning & Technology*, *18*(3), 173–192.	In contrast to the findings of Ruellot (2011), **Olson** reported that spectrograms were helpful for 26 intermediate-low proficiency learners of L2 Spanish (L1 English) to distinguish the stop consonants /b, d, g/ from their intervocalic variants, which are produced more like approximants. These variants are distinguishable on a spectrogram by the relative darkness of the image versus the silence of a stop's closure period. Findings revealed significant improvement and more native-like productions of the target sounds for those who had received visual feedback.	**B1**
2015	Chun, D. M., Jiang, Y., Meyr, J., & Yang, R. (2015). Acquisition of L2 Mandarin Chinese tones with learner-created tone visualizations. *Journal of Second Language Pronunciation*, *1*(1), 86–114.	Following the successful outcome of Wang (2012), first-year learners of Mandarin were taught to create pitch displays of their productions of bisyllabic words using *Praat* and compare them to a gender-matched native-speaker example. Training involved 20–25 minutes per week during class time over nine weeks, focusing on one tone per week. Two tones (1 and 4) showed substantial improvement, and learners offered positive feedback on their training.	**B1**
2015	Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *Journal of the Acoustical Society of America 138*(2), 817–832.	Several studies investigated production training with visual articulatory feedback such as ultrasound showing tongue position (e.g., Bliss et al., 2018; Engwall, 2012; Kocjančič Antolík, 2019). The focus of **Kartushina et al.** involved indirect articulatory feedback training on the production of non-native Danish vowels by native French speakers with no prior experience with the language. Participants received one hour of training for each of four vowels /e/ - /ɛ/, /y/ - /ø/. Feedback involved an immediate analysis of the acoustic properties of the vowels that participants had produced including visual feedback showing tongue height (e.g., a more open mouth tongue height is evident in a higher *F*1) and front-back position (e.g., a more back articulation has a lower *F*2 value). Post-training productions were closer to the Danish target vowels. In addition, production improvement transferred to improved perceptual discrimination of the vowels.	**B1**
2015	Katz, W. F., & Mehta, S. (2015). Visual feedback of tongue movement for novel speech sound learning. *Frontiers in Human Neuroscience*, *9*, 612.	Building on Katz et al. (2014), an interactive 3D tongue visualization system based on electromagnetic articulography was used to train native speakers of American English to produce a novel speech sound: a voiced palatal stop (requiring a closure between the tongue tip and hard palate). Training involved viewing their own real-time tongue movements using the 3D model, which produced a rapid gain in accuracy.	**B2**

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

2015	Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO*, *32*(1), 1–25.	This study reported significant improvement in the production of the French front rounded vowel /y/ by a group of elementary-level adult learners of French who used ASR-based pronunciation instruction on a mobile device. Activities, consisting of reading aloud target words and phrases, were accompanied by immediate orthographic feedback from an ASR dictation system. Results revealed significant improvement for learners using the ASR system versus those who practiced with a teacher and received oral feedback or participated in conversations with the teacher with no specific focus or feedback. The ASR group provided positive evaluation of its use. None of the groups improved in perception. The authors suggested the use of this technology as a complement to classroom instruction.	**B4**
2015	Morett, L. M., & Chung, L-Y., (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, *30*(3), 347–353.	This study explored whether hand gestures could facilitate English speakers' ability to discriminate between Mandarin words differing only in lexical tone. English monolinguals with no knowledge of Mandarin were assigned to one of three learning conditions in which the video of a Mandarin speaker varied according to gesture use as follows: pitch gesture (hand motions conveyed pitch contours), semantic gesture (hand motions conveyed word meaning), and no gesture. A pretest-learning-posttest sequence was used in which participants were instructed to view a series of video clips, and after each one, they repeated the word the speaker said and its English translation aloud while re-enacting any action they saw. Tone identification accuracy increased significantly from pre- to posttest in the pitch gesture condition and the no gesture condition, but not in the semantic gesture condition. This finding is consistent with Hirata and Kelly (2010) which found that hand gestures did not enhance perception of vowel length in Japanese by English listeners compared to an audio-only condition although speakers' lip movements were facilitative.	**C**
2015	Patten, I., & Edmonds, L. A. (2015). Effect of training Japanese L1 speakers in the production of American English /r/ using spectrographic visual feedback. *Computer Assisted Language Learning*, *28*(3), 241–259.	Ruellot (2011) and Olson (2014) used spectrographic displays as visual feedback in training for learners of L2 French vowels and L2 Spanish intervocalic approximants respectively. In **Patten and Edmonds'** study, two L1 Japanese participants produced isolated words involving American English /r/ in initial, medial, and final positions with a range of adjacent vowels while viewing a real-time spectrographic display to encourage a lowered third formant frequency (i.e., a key acoustic feature of /r/). Delayed visual feedback was then implemented to promote self-monitoring. Improvement was noted on trained and untrained isolated words, and those produced in a read passage. Perception of the /r/ - /l/ distinction also improved as noted in other training studies (e.g. Okuno & Hardison, 2016; Thomson, 2011).	**B1**
2016	Hisanaga, S., Sekiyama, K., Igasaki, T., & Murayama, N. (2016). Language/culture modulates brain and gaze processes in audiovisual speech perception. *Scientific Reports*, *6*, 35265.	**Hisanaga et al. (2016)** recorded a speaker of English and a speaker of Japanese producing /ba/ and /ga/. English and Japanese speakers were presented with these stimuli in their L1 for a syllable-identification task in the following conditions: AV and A-only to collect ERP (event-related brain potentials) data, and AV only for eye-tracking data. ERP data indicated that English speakers processed AV speech more efficiently than A-only, whereas Japanese speakers showed the opposite pattern. Eye-tracking data revealed a gaze bias to the mouth for English speakers, especially	**A, D**

(*Continued*)

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

(*Continued*)

Year	References	Annotations	Theme
		before the audio onset, but not for Japanese. Findings were somewhat compatible with the eye-tracking data of Hardison and Inceoglu (2019), and consistent with the influence of linguistic/cultural backgrounds on gaze behavior and speech processing. English speakers tend to focus on a speaker's mouth in contrast to Japanese speakers who focus on the voice in speech tasks and the eye region to interpret facial expression. However, L1 Japanese speakers enrolled in English instruction attend to the speaker's mouth when experiencing the McGurk Effect and show visual enhancement in English speech processing Hardison, 1996*, 2003*, 2005*; Hardison, 2018a).	
2016	Inceoglu, S. (2016). Effects of perception training on L2 vowel perception and production. *Applied Psycholinguistics*, *37*(5), 1175–1199.	In a pretest-perception training-posttest design, L2 French learners (L1 English) were divided into three groups to improve their identification accuracy of three French nasal vowels: AV training (visual input from a native speaker's face), A-only training, and no training. Similar to Hardison (2003*), both training groups showed significant improvement; in contrast, the AV group in **Inceoglu's** study did not show a significant advantage in perceptual accuracy. However, production accuracy improved significantly more for the group that saw the speaker's face.	**A**
2016	Offerman, H. M., & Olson, D. J. (2016). Visual feedback and second language segmental production: The generalizability of pronunciation gains. *System*, *59*, 45–60.	Following other studies involving electronic displays of acoustic information as training feedback at the segmental level (e.g., Okuno & Hardison, 2016; Olson, 2014; Patten & Edmonds, 2015), **Offerman and Olson** used visual feedback for voice onset time (VOT) training involving the voiceless stops /p, t, k/ in word-initial position produced by L1 English learners of L2 Spanish. Spanish is characterized by a shorter lag VOT compared to English, often resulting in a noticeable foreign accent in L2 speech. Participants recorded stimuli using *Praat* and were guided in the analysis of the spectrograms and waveforms. Significant VOT improvement (i.e., more native-like) was found for both the more controlled stimuli (read-aloud carrier sentences and short story) and less controlled (picture naming task) for the experimental (vs. control) group.	**B1**
2016	Okuno, T., & Hardison, D. M. (2016). Perception-production link in L2 Japanese vowel duration: Training with technology. *Language Learning & Technology*, *20*(2), 61–80.	Based on learners' positive comments and perceptual accuracy improvement involving the use of waveforms to visualize duration difference between L2 Japanese singleton and geminate consonants (Motohashi-Saigo & Hardison, 2009*), waveform displays were chosen as visual feedback in **Okuno and Hardison's** L2 Japanese vowel training study. L1 English learners were divided into two training groups: AV (saw waveform displays) or A-only training to improve the perceptual accuracy of Japanese vowel duration. Both types of training provided significant improvement, with a higher rate for the AV group, and no improvement for the control group. Greater perceptual accuracy transferred to greater production accuracy.	**B1**

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

2016	van Doremalen, J., Boves, L., Colpaert, J., Cucchiarini, C., & Strik, H. (2016). Evaluating automatic speech recognition-based language learning systems: A case study. *Computer Assisted Language Learning*, *29*(4), 833–851.	The *DISCO* prototype system developed by Strik et al. (2012) was evaluated by 'domain experts' with L2 Dutch teaching experience (no experience with ASR systems) and by learners (range of L1s) at the A2 (basic) level of the Common European Framework of Reference. Learners worked with *DISCO* for 45 minutes and completed a questionnaire. Teachers recognized the advantage of such a system for more introverted learners and recommended incorporating different strategies for responding to different types of pronunciation errors. Learners commented that the exercises were helpful and enjoyable, rating the system 7.8 out of 10, but suggested a more focused approach to error correction.	**B3, 4**
2016	Wallace, L. (2016). Using Google Web Speech as a springboard for identifying personal pronunciation problems. In J. Levis, H. Le, I. Lucic, E. Simpson, & S. Vo (Eds.). *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference* (pp. 180–186). Dallas, TX. Amex, IA: Iowa State University.	Google Web Speech (*GWS)* is an ASR-based transcription tool that can help learners build self-monitoring skills and identify pronunciation weaknesses. After *GWS* transcribes the speech, learners look at the discrepancies between the transcriptions and what they had said; for example, **Wallace** reported that in one instance *GWS* transcribed 'The person page' when the speaker had said 'The percentage'. Learners can focus on different features such as stress placement, thought group division, pitch movement, etc., and practice until the transcription is closer to their production. **Wallace** points out caveats to its effectiveness: (a) a headset microphone is needed for clear input, and (b) the learner's accent must fit within the parameters of the ASR model. She further suggests that individuals whose L2 English speech is not heavily accented would benefit the most. McCrocklin (2019) investigated another ASR-based dictation program, which could also be a useful pedagogical complement to classroom instruction.	**B4**
2017	Hacking, J. F., Smith, B. L., & Johnson, E. M. (2017). Utilizing electropalatography to train palatalized versus unpalatalized consonant productions by native speakers of American English learning Russian. *Journal of Second Language Pronunciation*, *3*(1), 9–33.	Using electropalatography (EPG) training, Schmidt (2012) had found improved pronunciation of several consonantal contrasts for L1 Korean learners of English. **Hacking et al.** used EPG feedback to highlight the tongue-palate contact, which is an important feature in distinguishing palatalized (e.g., /t^{j} /) versus nonpalatalized (e.g., /t/) consonants in Russian. Ten learners read several carrier sentences containing words contrasting /t^{j}/-/t/ and /s^{j}/-/s/. Eight training sessions involved learners monitoring their own tongue placement by looking at visual EPG targets produced by native speakers and listening to audio files. Learners showed significant increases in the frequency of the second formant of the vowel preceding palatalized consonants, which is an important cue. Native listeners' ratings revealed only small improvements in identification accuracy of the sounds.	**B2**
2017	Venezia, J. H., Vaden, Jr., K. I., Rong, F., Maddox, D., Saberi, K. & G. Hickok (2017). Auditory, visual and audiovisual speech processing streams in superior temporal sulcus. *Frontiers in Human Neuroscience*, *11*(174).	The human superior temporal sulcus (STS), located in the temporal lobe of the brain, responds to visual and auditory information. Using an fMRI design, **Venezia et al.** measured activation in native speakers of English to a range of auditory and visual speech (A-only, V-only, and AV) and nonspeech stimuli with a focus on the patterns of activation within STS. Speech-specific activations arose in multisensory regions of the middle STS; abstract representations of visible facial gestures emerged in visual regions that immediately border the multisensory regions. The middle STS also exhibited preferential responses for speech versus nonspeech stimuli.	**D**

(*Continued*)

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

(*Continued*)

Year	References	Annotations	Theme
2018	Bliss, H., Abel, J., & Gick, B. (2018). Computer-assisted visual articulation feedback in L2 pronunciation instruction: A review. *Journal of Second Language Pronunciation*, *4*(1), 129–153.	**Bliss et al.** review computer-assisted visual displays in the form of direct feedback on articulation, such as ultrasound imaging (e.g. Gick et al., 2008*) to observe the position and movement of the tongue, and indirect feedback using displays of acoustic information such as pitch (e.g., Chun et al., 2015) or waveforms (e.g., Okuno & Hardison, 2016). Bliss et al. noted that ultrasound feedback is more informative for vowel articulations, laterals (/l/), and rhotics (/r/-like sounds).	**B2**
2018	Cucchiarini, C., & Strik, H. (2018). Automatic speech recognition for second language pronunciation training. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 556–569). London, UK: Routledge.	**Cucchiarini and Strik** provide a recent overview of developments in ASR within the context of L2 pronunciation. They point out that annotated corpora of native and learner speech are needed to develop ASR-based systems so they can be trained to recognize areas where learner utterances deviate from the target. The assessments of pronunciation quality can be used as a basis for providing feedback, which can include visual representations of articulations (Cucchiarini et al., 2009; Engwall, 2012; Van Doremalen et al., 2016). The authors emphasize that ASR-based systems should not be viewed as substitutes for pronunciation instruction by teachers.	**B4**
2018a	Hardison, D. M. (2018). Effects of contextual and visual cues on spoken language processing: Enhancing L2 perceptual salience through focused training. In S. M. Gass, P. Spinner, & J. Behney (Eds.), *Salience in second language acquisition* (pp. 201–220). New York, NY: Routledge.	Previous studies had shown that visual cues from a speaker's face contributed to segmental perceptual accuracy for a variety of populations including L2 learners (e.g., Hardison, 2003*). This study found that for L2 learners of English (L1 Japanese and Korean), AV (vs. A-only) training resulted in earlier identification of words presented in isolation and in sentence contexts. Both the temporal precedence of visible articulatory gestures (Munhall & Tohkura, 1998*; Navarra et al., 2010), their increased salience following training, and the presence of context served priming roles in reducing the initial cohort of word candidates in the recognition process. For L2 learners, visual cues and contextual cues had independent effects statistically in contrast to the native speakers for whom the variables showed a significant interaction; specifically, only 62% of a word was needed for identification when both types of cues were present.	**A**
2018b	Hardison, D. M. (2018). Visualizing the acoustic and gestural beats of emphasis in multimodal discourse: Theoretical and pedagogical implications. *Journal of Second Language Pronunciation*, *4*(2), 231–258.	Annotations from *Praat*, a phonetic analysis tool, and *ANVIL*, a video annotation tool were combined to provide a time-aligned display of visual (gestural) and acoustic beats in the natural speech of native and non-native teachers of English. Frame-by-frame analysis revealed several points of temporal convergence such as maximum brow raise and upright head position with pitch-accented vowels. The temporal interval between the apex (most extended position) of each beat gesture was fairly regular except for the lengthening that occurred around pitch-accented vowels. These polyrhythmic sequences (i.e., those with different rhythms for speech and gesture) were found to be perceptually salient highlighters of important information for students (see also Dimitrova et al., 2016*).	**B1, C**

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

2019	Hardison, D. M. (2019). *Relationships among gesture type, pitch, and vowel duration in the speech of native-speaking teachers of Japanese.* Manuscript in progress.	**Hardison** investigated (a) the temporal coordination of naturally occurring gestures (head nods and hand-arm movements) by three native-speaking classroom teachers of beginning-level Japanese in Japan and two speech phenomena: pitch movement and vowel duration, and (b) the influence of these gestures on perceptual accuracy by second-year L2 learners of Japanese (L1 English). Videorecordings were analyzed with *Praat* and *ANVIL* (video annotation tool) allowing temporal integration of videorecorded gestures with the pitch contour and waveform (see also Hardison, 2018b). A significantly greater number of head nods occurred with a long (vs. short) vowel for all teachers. The apex of the head movement coincided with the peak of the syllable containing the long vowel and pitch contour. Fewer hand gestures were used but tended to occur with short vowels (see Hirata et al., 2014). In contrast to Hirata and Kelly (2010), learners' perceptual accuracy in identifying vowel duration was greatest when head movement and facial cues were present, followed by facial cues (no head movement), and then the A-only and V-only conditions.	**C**
2019	Hardison, D. M., & Inceoglu, S. (2019). *L1 and L2 auditory-visual speech perception: Using eye tracking to investigate effects of task difficulty.* Manuscript in preparation.	Cues from talkers' faces significantly enhance speech processing (e.g., Hardison, 2003, 2005, see Introduction Hardison, 2018a; Inceoglu, 2016; Yi et al., 2013). **Hardison and Inceoglu** used eye tracking to investigate where and when participants looked on a speaker's face while processing speech in L1 English and L2 French under different conditions: AV, AVn (AV with noise added), V-only. Thirty-two participants (L1 English) viewed one English and one French native speaker each producing stimuli involving minimal triplets, differing in jaw height for English front vowels and lip rounding for French nasal vowels. Areas of eye-gaze interest were the forehead, each eye, nose, mouth, and lower jaw. For both languages, fixations occurred (a) to the forehead infrequently but early; (b) to the nose early across modalities; and (c) to the mouth earlier in V and AVn conditions (vs. AV). Fixation durations increased to the mouth and decreased to the eyes with degraded or no audio. Incremental examination of heat maps and gaze patterns revealed fixations were often made centrally to the nose with strategic shifts of attention to other areas, especially the mouth (the most informative area) with the slightest articulation-related movement.	**A**
2019	Inceoglu, S., & Gnevsheva, K. (2020). *Ultrasound imaging in the foreign language classroom: Outcomes, challenges, and students' perceptions.* In O. Kang, S. Staples, K. Yaw, & K. Hirschi (Eds.), *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching Conference* (pp. 115–126). Northern Arizona University, September 2019. Ames, IA: Iowa State University.	This research extended the use of ultrasound technology (see Bliss et al., 2018) to a language classroom setting. L2 French learners (L1 English) received two lessons involving ultrasound visual feedback on their articulation of the vowel contrasts [y]-[u] or [e]-[ɛ]. Some improvement was noted in the production of [y] in word lists but not in a reading passage. Participants offered very positive comments on the use of ultrasound technology.	**B2**

(*Continued*)

https://doi.org/10.1017/S0261444820000592 Published online by Cambridge University Press

(*Continued*)

Year	References	Annotations	Theme
2019	Kocjančič Antolík, T., Pillot-Loiseau, C., & Kamiyama, T. (2019). The effectiveness of real-time ultrasound visual feedback on tongue movements in L2 pronunciation training. *Journal of Second Language Pronunciation*, *5*(1), 72–97.	Based on the positive outcome of the pilot study by Gick et al. (2008*) involving ultrasound feedback in L2 pronunciation training, L1 Japanese learners of French received three individual 45-minute lessons involving production of the French vowel [y]-[u] contrast using ultrasound feedback. Results showed improvement in their production of the French vowels, and in the contrast between the French vowels and the high back unrounded Japanese vowel.	**B2**
2019	McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation*, *5*(1), 98–118.	ASR-based technology has been explored in some studies, especially for L2 Dutch (Cucchiarini et al., 2009; Van Doremalen et al., 2016), as a resource for learners to monitor their speech or help detect errors; however, this technology is less accessible. ASR-based dictation programs (e.g., *Windows Speech Recognition*) are more accessible (e.g., Wallace, 2016). In a pretest-posttest design, **McCrocklin** found that L2 English learners who received both face-to-face instruction and practice using the dictation program showed significant pronunciation improvement as did a group receiving only face-to-face instruction. While the groups did not differ significantly, results suggested that dictation programs may be a useful pedagogical complement to classroom instruction.	**B4**
2019	Zheng, Y., & Samuel, A. G. (2019). How much do visual cues help listeners in perceiving accented speech? *Applied Psycholinguistics*, *40*(1), 93–109.	Yi et al. (2013) found that visual cues benefited perception of speech produced by a native versus a non-native speaker. **Zheng and Samuel** used a lexical decision task to explore the intelligibility of speech produced by native English speakers and two non-native English speakers who differed in the strength of their accent. Stimuli were relatively frequent words plus nonwords. Two versions of each videorecorded token were created: (a) one with a speaker far away, and (b) one focused on the speaker's head. The audio was the same. Accuracy was greater for L1 English listeners when they could see a speaker's lip movements at a closer distance, and this effect was slightly stronger for recognition of nonwords versus words, and stimuli produced with a stronger accent. There was no apparent influence of listeners' prior experience with Mandarin-accented speech.	**A**