Proceedings of the International Symposium on the Acquisition of Second Language Speech
Concordia Working Papers in Applied Linguistics, 5, 2014 © 2014 COPAL

Effect of Multimodal Training on the Perception of French Nasal Vowels

Solene Inceoglu
Michigan State University

Abstract

This study investigates the effects of multimodal training and the contribution of facial cues on the perception of French nasal vowels by learners of French. Sixty American English learners of French participated in the study. Forty were assigned to an experimental training group (either Audio-visual or Audio-only training) and 20 served as controls. All participants completed a perception pretest and a perception posttest presented within three modalities and with two counterbalanced orders: AV, A, V or A, AV, V. During the three weeks between the pre- and posttests, the AV and A groups received six sessions of perception training with immediate feedback. Results show significant improvement for the training groups in the perception of the three vowels at the posttest and generalization test, but the AV training was not better than the A training. As for the vowels, the greatest improvement was obtained for [ɛ̃].

Keywords: multimodal training, audiovisual perception, vowel perception, French nasal vowels.

Face to face interaction often involves the simultaneous perception of the speaker's voice and facial cues, such as lip movements, making speech perception a multimodal experience (Rosenblum, 2005). Such facial cues have been shown to contribute to comprehension in noisy environments (Benoît, Mohamadi, & Kandel, 1994) or when people suffer from impaired hearing ability (Owens & Blazek, 1985). One of the most famous examples demonstrating the contribution of both oral and visual information on speech perception is probably the "McGurk effect" (McGurk & MacDonald, 1976), where, among other stimuli, an audio /ba/ was dubbed onto a visual /ga/ and was perceived by native speakers of English as a /da/. The mere fact of observing lip movements, without having access to any speech sounds, was shown to activate the auditory cortex, reinforcing the idea that "seen speech" influences "heard speech" (Calvert et al., 1997). Evidence points out that visual information alone allows discrimination between pairs of languages (for English/French, see Weikum et al., 2007). In addition, second language (L2) speech research has also noted the importance of variables such as the perceiver's familiarity with the talker, the linguistic background or the native language (L1), the phonetic environment, the native or nonnative status of the observer's perceptual categories, or the proficiency in visual L2 speech perception. Werker, Frost, and McGurk (1992) found a positive correlation between proficiency in visual L2 speech perception and the overall L2 proficiency by investigating French native speakers with varying levels of English and their perception of incongruent audio-visual English fricatives. The majority of research on visual and auditory input in L2 speech processing has investigated consonant sounds. These studies often looked at the contrast between problematic L2 sounds, which are usually absent in certain L1s, such as the English /r/ and /l/ (e.g. Hardison, 2003; Lively, Logan, & Pisoni, 1993), English fricatives (Wang, Behne, & Jiang, 2008) or the /b-v/ contrast in English (Hazan et al., 2006). For instance, because French does not have interdental fricatives, French beginner and intermediate learners of English cannot efficiently take advantage of the visible information in order to accurately identify the interdental fricative (Werker et al., 1992).

Visual intelligibility of vowels has also received some attention, despite the fact that vowels might appear less visually salient than consonants since the major determinant of the acoustical structure of vowels is the position of the body of the tongue (Stevens & House, 1955). However, as Summerfield, (1991) notes, "under optimal conditions, all English vowels are visibly distinct" (p. 119). This claim can be extended to French vowels

since there is a strong correlation between the height of the tongue in the mouth and the vertical separation of the lips. The visibility of the teeth and the spreading/rounding dimension are also important factors when considering lipreading. Research on audio-visual intelligibility of vowels has looked at various languages, taking into consideration perceivers from different L1s and various types of vowel distinctions. For example, Navarra and Soto-Faraco (2007) found that, contrary to bilinguals with Catalan as their dominant language, Spanish-Catalan bilinguals with Spanish as their dominant language cannot distinguish the Catalan vowel contrast [ɛ - e] in audio-only condition. However, when visual articulatory information is added, both groups can perceive the contrast, and when only visual information (e.g., no sound) is presented, none of the group can discriminate the two phonemes. In a study investigating whether multimodal input, namely information such as lip movements and hand gestures, helps improve native English speakers' ability to perceive Japanese vowel length contrasts, Hirata and Kelly (2010) found that, after four sessions training participants to distinguish long and short vowels, all experimental groups improved from pre- to posttest, but that the improvement was greatest when mouth movements accompanied the auditory training, as opposed to hand gestures. Moreover, the results suggested that the participants in the AV condition improved more than those in the A condition.

What distinguishes nasal vowels from oral vowels is that they are produced with a lowering of the velum so that air escapes both through nose as well as the mouth. It is often argued that AE learners of French encounter difficulty with the perception and production of French nasal vowels because the English phonemic inventory does not possess nasal vowels. English does, however, possess vowels which are similar to French nasal vowels, but their nasalization occurs because of the phonetic environment (preceding a nasal consonant because of regressive assimilation), rather than as a distinction between minimal pairs (Valdman, 1961). While some French dialects distinguish four different nasal vowels [ɑ̃ - ɔ̃ - ɛ̃ - œ̃], in Parisian French the [œ̃] has been progressively replaced by the [ɛ̃] so that there are now for the most part only three nasal vowels (Walter, 1977). The current study investigates the perception of the three Parisian French nasal vowels: [ɔ̃], [ɛ̃] and [ɑ̃]. The vowel [ɑ̃] is articulated with a protrusion and a narrowing of the interlabial gap, making it somewhat a rounded vowel. Similarly, the back nasal vowel [ɔ̃] is "hyperrounded" and comparable to the oral mid-close

[o]. Finally, [ɛ̃] is an unrounded vowel. From an articulatory perspective, the three nasal vowels differ in terms of rounding, but also in terms of height, namely by the distance between the jaws, between the tongue and the palate, and also between the upper and lower lips, the latter being the most visually salient cue.

The present study attempts to investigate the effects of multimodal training and the contribution of facial cues in the perception of French nasal vowels by L2 learners of French, and more specifically to answer the following research questions:

1. How does perception accuracy vary according to the vowel and the modality?
2. Is there better perception performance with AV training than with A training?
3. Is training generalizable to novel stimuli?

METHOD

Participants

Sixty L1 American English (age 18-24, mean = 20) intermediate-level learners of French participated in the study. Forty were randomly assigned to one of the two experimental groups (AV or A training) and 20 served as controls. Participants were asked to fill out background questionnaires reporting information such as participants' possible additional L1s, stay-abroad, amount of visual input in French (e.g., movies, conversation with native speakers), age, additional L2s, and background in phonetics and linguistics. None of the participants had training in lipreading and all reported good vision and no hearing disorders. Participants were paid $60 or $20 for their time and all received 10 extra-credits for their French classes. The data of one participant from the control group was excluded from the analysis because she did not complete the tasks properly.

Stimuli

Stimuli were triads of the three French Parisian vowels [ɑ̃ - ɔ̃ - ɛ̃] in various consonantal contexts. A total of 396 stimuli was used for the pretest/posttest, training and generalization: 324 #CVC# stimuli where C was one of the following consonants [p-t-k-b-d-g-s-z-f-v-ʒ-ʃ], for example [pɔ̃t], 36 initial consonant clusters #CcVC# where the cluster was [dʁ] and

the final consonant one of [p-t-k-b-d-g-s-z-f-v-ʒ-ʃ], for example, [dʁɔ̃t], and 36 final consonant cluster #CVCc# where the first consonant was one of the following [p-t-k-b-d-g-s-z-f-v-ʒ-ʃ] and the final cluster [dʁ], for example, [kɔ̃dʁ]. The stimuli with the consonantal clusters were only used for the generalization test.

Previous perceptual studies have used nonsense words (Iverson, Pinet, & Evans, 2011; Levy & Strange, 2008) or real words (Hardison, 2003), however in an experiment with triads of French nasal vowels it is impossible to use only nonsense words or only real words. Nevertheless, because the purpose of the experiment was to discriminate sounds, the task for the listeners was to focus only on specific sounds under test and it is therefore improbable that their potential background vocabulary knowledge influenced their choices. In addition, previous studies (see Flege, Takagi, & Mann, 1995) showed no correlation between word familiarity and identification scores in speech perception experiments.

Procedure

Stimuli Recording. A female native speaker of French in her early thirties was videorecorded in a quiet research room. She was instructed to read lists of nonsense words containing nasal vowels in the most natural fashion with the aim of avoiding hyperarticulation. The camera captured a full-sized image of the speaker's head and her lower jaw drop was fully visible. The recording was used for both the AV and the A conditions in order to prevent any possible aural variations for each token across the two conditions.

Pretest and Posttest. Participants were presented with 108 stimuli in each of the three modalities: AV, A and V, yielding a total of 324 responses per participant. The number of stimuli was not higher to ensure that participants would not get tired and to keep the whole duration of the pretest session under one hour. The list of stimuli was comprised of 108 items with a balanced distribution of vowel, and place and manner of articulation for the consonants. The order of items was different for each modality and was counterbalanced across participants. Two orders were used for the presentation of the three modalities: AV, A, V or A, AV, V, and the two orders were also counterbalanced across participants. The experiments were conducted on a MAC computer using SuperLab. Participants heard the stimulus and were asked to select one of the three sound representing on the screen. <on> always appeared on the left, <an>

in the middle, and <un> on the right. Participants had 4 seconds to click on one of the three options before presentation of the next stimulus. No feedback was given. Before the beginning of the experiment, a practice task was administered to familiarize participants with the procedure, and to ensure that the volume was adequate. Following the post-test, a generalization test with 108 novel stimuli was administered.

Training. Training consisted of six sessions completed within a two-week period. The training sessions followed the same procedures as for the pretest, except that participants received feedback. A total of 178 stimuli, distinct from the pretest, was used for each session. The order of presentation was randomized across participants and across sessions. Each training sessions lasted about 30 minutes.

RESULTS

The first research question investigated learners' perception accuracy according to vowel and modality. The results are presented in Figure 1 and show that two different hierarchies of identification were found. In the A and AV modalities, L2 learners were better at correctly identifying the vowel [ɔ̃], followed by [ɑ̃] and [ɛ̃]. In the V modalities, [ɔ̃] was still the most intelligible vowel, but [ɛ̃] was found more intelligible than [ɑ̃]. A repeated-measures ANOVA with Vowel ([ɑ̃ - ɔ̃ - ɛ̃]) and Modality (A, AV, V) as within-subjects variables was performed. Results indicated that the main effect of Vowel, $F(2, 116) = 19.44$, $p < .0001$, and Modality ($F(2, 116) = 166.12$, $p < .0001$) were significant. Post-hoc pairwise comparisons revealed statistically significant differences between each vowel within each of the three modalities, between AV and V for all vowels, between A and V for [ɔ̃] and [ɑ̃], and between A and AV for [ɑ̃].

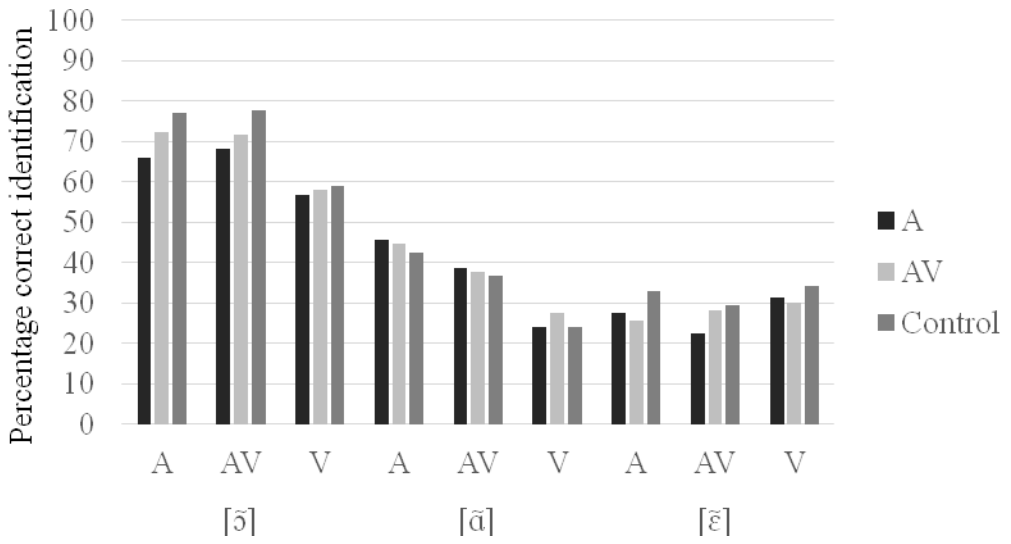


Figure 1. Percentage of correct identification score at the pretest for each vowel and modality.

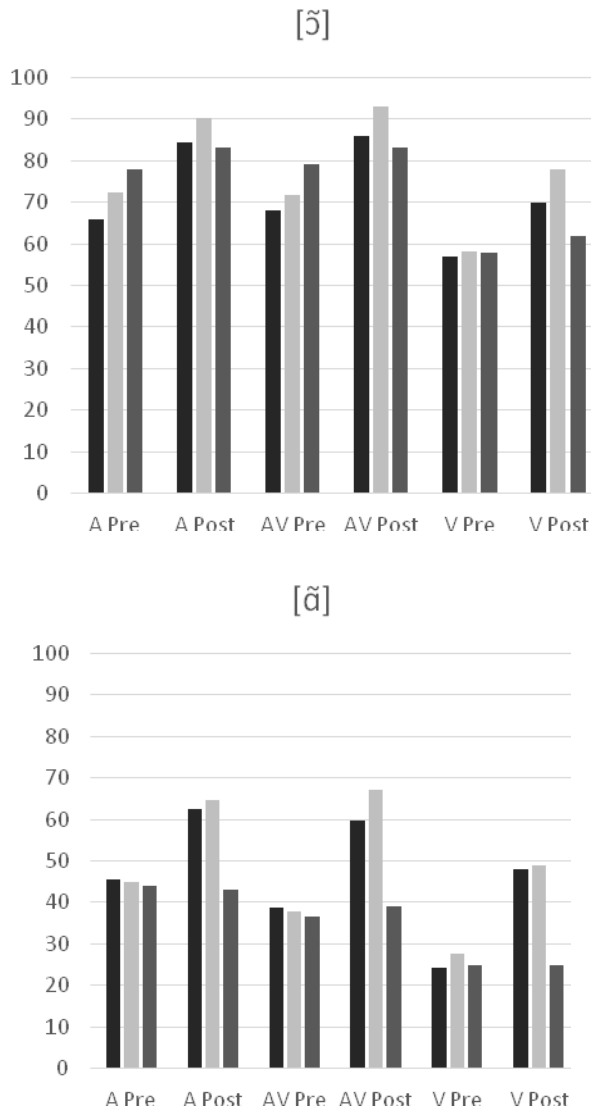
To assess the effectiveness of training condition, a repeated-measures ANOVA was conducted on the pretest and posttest identification scores for the two training groups separately. The variables were Time (pretest, posttest), Modality (A, AV, V) and Vowel (ɑ̃ - ɔ̃ - ɛ̃). Figure 2 illustrates the scores for the two training groups and the control group. The results of the control group were not significantly different from the pretest to the posttest, $F(1, 18) = 75.79$, $p = .18$.

For the AV training group, there was a significant main effects of Time, $F(1, 19) = 55.56$, $p < .0001$, Modality, $F(2, 38) = 18.38$, $p < .0001$, and Vowel, $F(2, 38) = 40.75$, $p < .0001$. The interaction of Time × Modality, $F(2, 38) = 2.84$, $p < .070$, indicates that no modality of presentation improved significantly more than others. Comparisons of posttest performances showed an increase of about 29% in the A modality, 34% in the AV modality and 25% in the V modality. The interaction Time × Vowel was however significant $F(2, 38) = 12.42$, $p < .0001$, with a performance increase of 44% for [ɛ̃], 23% for [ɑ̃], and 20% for [ɔ̃].

For the A training group, there was a significant main effects of Time, $F(1, 19) = 63.42$, $p < .0001$, Modality, $F(2, 38) = 23.07$, $p < .0001$, and Vowel, $F(2, 38) = 51.90$, $p < .0001$. The interaction of Time × Modality, $F(2, 38) = 48.51$, $p = .08$, indicates that, similarly to the AV group, none of the modality improved more than the others. Comparisons of posttest performances showed an increase of about 28% in the A modality, 30% in the AV modality and 23% in the V modality. The interaction Time × Vowel

was also significant $F(2, 38) = 17.72$, $p < .0001$, with a performance increase of 45% for [ɛ̃], 20% for [ɑ̃], and 16% for [ɔ̃].

A series of repeated-measures ANOVAS with Time as within-subjects factor and Training type as between-subjects factor were run to compare the effect of training condition on posttest performances for the three modalities. Time was a statistically significant factor (at $p < .0001$) for the three modalities, but the interaction Time × Training was not statistically significant for the A condition ($p = .86$), AV condition ($p = .58$), and V condition ($p = .81$), indicating that neither of the training type provided greater improvement in perceptual accuracy.



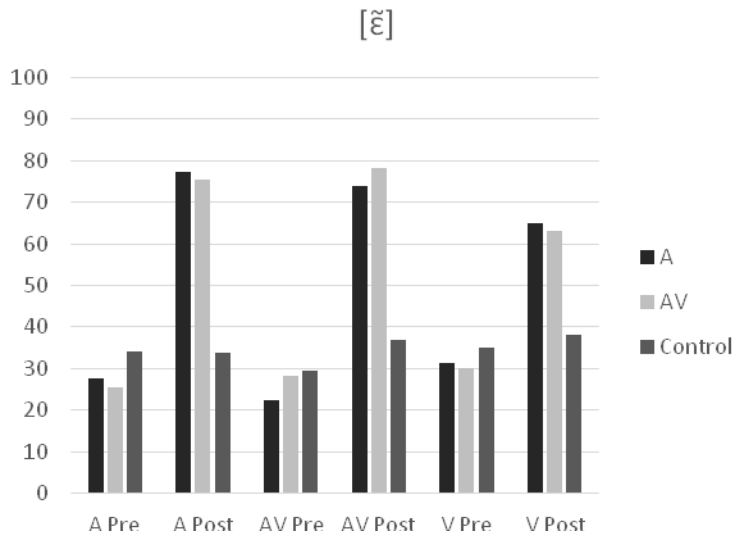


Figure 2. Percentage of correct identification score according to vowel, modality and pre/posttest.

Figure 3 illustrates the percentage of correct identification score for the generalization test given right after the posttest and investigating learners' ability to generalize to novel stimuli. Again, the results of the control group were not significantly different from the posttest to the generalization test, $F(1, 18) = .399$, $p = .53$.

For the AV group, mean identification accuracy at the generalization test was significantly lower than at the posttest in the A and AV modalities, but not in the V modality. Analysis of the results per vowel revealed that performance decreased significantly from the posttest to the generalization test for [ɛ̃] in the A and AV modalities, but improved for [ɔ̃] in the V modality.

For the A group, mean identification accuracy was similar from the posttest to the generalization test in the A and AV modalities, but there was a significant improvement at the generalization test in the V modality. Analysis of the results per vowel showed that performance decreased for [ɛ̃] in the A modality.

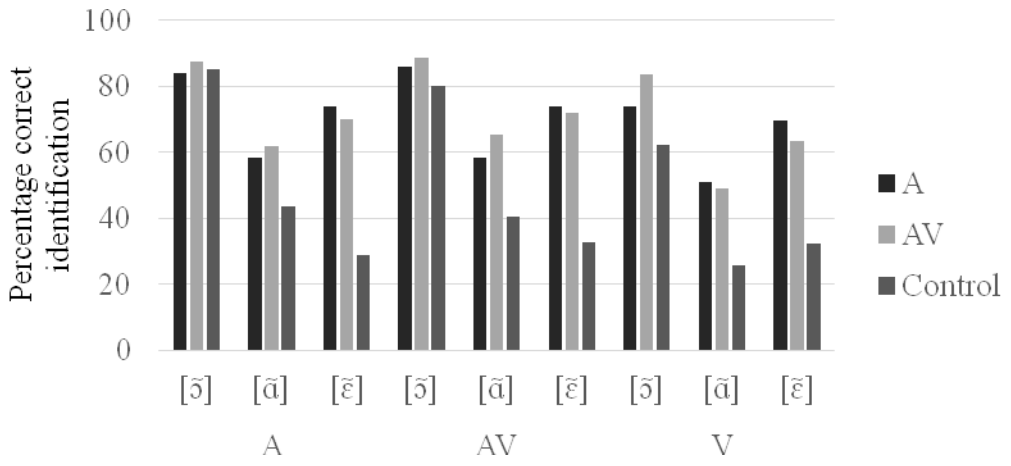


Figure 3. Percentage of correct identification score for the generalization test.

DISCUSSION

The purpose of this study was to compare A and AV trainings on the perception of French nasal vowels. Results of the pretest showed that in the A and AV modalities, the hierarchy of intelligibility by L2 learners was [ɔ̃] > [ɑ̃] > [ɛ̃], but that in the V modality, where no audio cue was present, [ɛ̃] was more intelligible than [ɑ̃]. Taking into account the importance of labial gesture, the fact that [ɑ̃] received the lowest accuracy scores is probably due to its intermediate position on the continuum of labiality and therefore to its lack of visual saliency as opposed to the hyperrounded [ɔ̃] and the unrounded [ɛ̃].

The performance of both training groups significantly improved from the pretest to the posttest, contrary to the control group, demonstrating that training indeed helped improve perceptual accuracy. Training was particularly beneficial for [ɛ̃] and results of the posttest revealed that, although the control group retained that same hierarchy of intelligibility as the pretest (i.e. [ɔ̃] > [ɑ̃] > [ɛ̃]), results for both training group now yielded a [ɔ̃] > [ɛ̃] > [ɑ̃] hierarchy in all modalities. Identification accuracy was also successfully generalizable to novel stimuli, replicating the effects of training from previous studies (Hardison, 2003).

Nevertheless, it is important to note that contrary to previous studies on audio-visual speech perception, in the current study, training two modalities (A and V) simultaneously was not superior in improving perceptual accuracy to training only one. In particular, learners trained

audiovisually did not improve their perception of the vowels in the V condition significantly more than those trained auditorily. This might be explained by a possible effect of overwhelming perceptual information due to the consonantal context. Further analysis will investigate the role of the preceding and following consonant on the audio and visual intelligibility of the vowels. In addition, difficulties in categorizing the L2 vowels may be caused by a less accurate processing or by less well-defined (but existing) representations, but improvement at the posttest showed that L2 learners do not lack these L2 phonemic categories. Finally, studies in L2 acquisition report strong evidence of individual variability, and so even when learners receive the same exposure to the L2. There is also great variability in learners' lipreading skills (Summerfield, 1992) and it is possible that some learners benefited more from the training condition they were assigned to than others. Because of variability in learning styles, it is also plausible that the visual cues in the AV modality acted as distractors for aural learners. Further analyses will look at individual results to see how variability affected the general patterns of the study.

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), 1195–203.

Calvert, G. a, Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596.

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English /I/ and /l/ accurately. *Language and Speech*, *38*(1), 25–55.

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

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America*, *119*(3), 1740–1751.

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*, 298–310.

Iverson, P., Pinet, M., & Evans, B. G. (2011). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 1–16.

Levy, E. S., & Strange, W. (2008). Perception of French vowels by American English adults with and without French language experience. *Journal of Phonetics*, *36*(1), 141–157.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in

learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3), 1242–1255.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12.

Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal Of Speech And Hearing Research*, *28*(3), 381–393.

Rosenblum, L. D. (2005). Primacy of multimodal speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 51–78). Malden, MA: Blackwell.

Stevens, K., & House, A. (1955). Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America*, *27*(3), 484–493.

Summerfield, Q. (1991). Visual perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 117–137). Hillsdale, NJ: Erlbaum.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions: Biological Sciences*, *335*(1273), 71–78.

Valdman, A. (1961). Teaching the French vowels. *The Modern Language Journal*, *45*(6), 257–262.

Walter, H. (1977). *La phonologie du français.* Paris, France: Presses Universitaires de France.

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, *124*(3), 1716–1726.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science New York NY*, *316*(5828), 1159.

Werker, J. F., Frost, P. E., & McGurk, H. (1992). La langue et les lèvres: Cross-language influences on bimodal speech perception. *Canadian Journal of Psychology*, *46*(4), 551–568.