Lexically guided retuning of visual phonetic categories

Patrick van der Zande[a)]
Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 A.H. Nijmegen, The Netherlands

Alexandra Jesse
Department of Psychology, University of Massachusetts Amherst, 135 Hicks Way, Amherst, Massachusetts 01003

Anne Cutler
MARCS Institute, University of Western Sydney, Penrith South, New South Wales 2751, Australia

(Received 5 September 2012; revised 15 January 2013; accepted 7 May 2013)

Listeners retune the boundaries between phonetic categories to adjust to individual speakers' productions. Lexical information, for example, indicates what an unusual sound is supposed to be, and boundary retuning then enables the speaker's sound to be included in the appropriate auditory phonetic category. In this study, it was investigated whether lexical knowledge that is known to guide the retuning of auditory phonetic categories, can also retune visual phonetic categories. In Experiment 1, exposure to a visual idiosyncrasy in ambiguous audiovisually presented target words in a lexical decision task indeed resulted in retuning of the visual category boundary based on the disambiguating lexical context. In Experiment 2 it was tested whether lexical information retunes visual categories directly, or indirectly through the generalization from retuned auditory phonetic categories. Here, participants were exposed to auditory-only versions of the same ambiguous target words as in Experiment 1. Auditory phonetic categories were retuned by lexical knowledge, but no shifts were observed for the visual phonetic categories. Lexical knowledge can therefore guide retuning of visual phonetic categories, but lexically guided retuning of auditory phonetic categories is not generalized to visual categories. Rather, listeners adjust auditory and visual phonetic categories to talker idiosyncrasies separately. © *2013 Acoustical Society of America.*
[http://dx.doi.org/10.1121/1.4807814]

PACS number(s): 43.71.Bp, 43.71.Es [MSV] Pages: 562–571

I. INTRODUCTION

In everyday communication, listeners encounter a variety of talkers, and all of them may pronounce the sounds of their native language in their own specific, idiosyncratic way. Such variation between speakers can arise from physiological differences (Laver and Trudgill, 1979), or because speakers have different dialectal and sociological backgrounds (Foulkes and Docherty, 2006). Given proper disambiguating information, however, listeners quickly and effectively adjust phonetic category boundaries to incorporate a speaker's idiosyncratic realizations of sounds into the correct phonetic categories (Bertelson *et al.*, 2003; Norris *et al.*, 2003; Baart and Vroomen, 2010; Jesse and McQueen, 2011). In face-to-face communication, listeners also make use of visual information about their interlocutors' articulation, and in doing so they draw on visually defined categories for individual phonemes (Van Son *et al.*, 1994; Massaro, 1998). Idiosyncratic articulations may also require the retuning of these visual phonetic categories. Simultaneously presented auditory information that disambiguates the sound can guide such retuning (Baart and Vroomen, 2010). Suppose, however, that an idiosyncratic articulation results in a sound being simultaneously both visually and auditorily ambiguous. In that case, the listener may still use lexical knowledge to guide retuning. But, is one retuning operation then needed, or two? We investigate here whether lexical knowledge (known at least to retune auditory category boundaries; Norris *et al.*, 2003) can lead to a retuning of visual phonetic categories in the absence of explicit auditory disambiguation. We further test whether retuning of visual phonetic categories can occur through generalization across modalities. Can retuning of auditory phonetic categories on the basis of lexical information also result in shifts of visual category boundaries?

Norris and colleagues (2003) showed that knowledge about the words of listeners' native language not only disambiguates idiosyncratic sounds, but also results in shifts in listeners' auditory phonetic category boundaries. Dutch listeners were presented with either /s/-final words such as *radijs* "radish," or /f/-final words such as *olijf* "olive" where the final fricative sound was replaced with an ambiguous sound between /s/ and /f/. Despite this alteration, listeners accepted these words in lexical decision. In a subsequent categorization task, listeners who had been exposed to the ambiguous sound in words normally ending in /s/ categorized more sounds from an /s/-/f/ continuum as /s/ than listeners exposed to the same sound in words normally ending in /f/. Thus, reference to existing knowledge allows category boundaries to be rapidly adjusted to incorporate an ambiguous sound into the appropriate phonetic category. This lexically guided retuning can be speaker-specific (Eisner and McQueen, 2005), and is stable in that its effects last at least

a)Author to whom correspondence should be addressed. Electronic mail: patrick.vanderzande@mpi.nl

 0001-4966/2013/134(1)/562/10/$30.00

for 24 h (Eisner and McQueen, 2006). Besides for fricatives, as in these studies, this retuning has been demonstrated for stop consonants (Kraljic and Samuel, 2006) and liquids (Scharenborg *et al*., 2011), as well as for lexical tone in Mandarin (Mitterer *et al*., 2011).

Importantly, retuning facilitates speech recognition in any situation where a similar idiosyncrasy is encountered. The effect of lexically guided retuning for auditory phonetic categories generalizes across word-internal positions and also generalizes to novel words (McQueen *et al*., 2006a; Sjerps and McQueen, 2010; Jesse and McQueen, 2011; Mitterer *et al*., 2011). Listeners who were exposed to an ambiguous fricative between /f/ and /s/ in word-final position showed, for example, boundary shifts in line with their exposure even when the ambiguous fricative occurred in word-initial position (Jesse and McQueen, 2011). In another study, listeners performed a cross-modal priming task at test that included auditory primes ending in the ambiguous fricative. The ambiguous auditory primes, e.g., /naɪ?/, could be interpreted as either an /f/-final word ("knife") or an /s/-final word ("nice"). The pattern of priming from these ambiguous auditory tokens revealed that they were interpreted by listeners in line with the listeners' prior exposure (McQueen *et al*., 2006a; Sjerps and McQueen, 2010). Phonetic retuning thus allows listeners to deal with the considerable variability that speakers show in their pronunciation of the sounds of their native language.

Communication is not a purely auditory phenomenon, however, and spoken interaction also provides visual information, for instance, concerning articulatory movements. In face-to-face communication, listeners automatically combine information obtained from hearing and seeing a speaker (Massaro, 1987; 1998). Visual speech affects identification even when listeners are instructed to disregard talkers' mouth movements (McGurk and MacDonald, 1976; Massaro and Cohen, 1983). This use of visual speech information is typically beneficial to the listener as it improves the intelligibility of a speaker significantly (e.g., Macleod and Summerfield, 1987; Reisberg *et al*., 1987; Jesse *et al*., 2000/2001; Helfer and Freyman, 2005; Spehar *et al*., 2008). Bimodal speech perception is especially useful when the input in one modality is difficult to interpret (Sumby and Pollack, 1954). The information provided by the two modalities is redundant but also complementary in that phonetic features that are difficult to distinguish in one modality are often more easily distinguished in the other modality (Walden *et al*., 1974; Summerfield, 1987; Grant *et al*., 1998; Jesse and Massaro, 2010). Because of this, audiovisual speech recognition performance often exceeds the simple addition of auditory-only and visual-only performances (Massaro and Cohen, 1983; Massaro and Friedman, 1990). The benefit of bimodal speech perception over unimodal perception decreases, for example, with increased redundancy between the information from the two modalities (Grant *et al*., 1998).

The influence of visual speech input goes beyond simple facilitation of recognition through disambiguation. Like lexical information, visual speech input guides the retuning of auditory phonetic categories (Bertelson *et al*., 2003). Simultaneously presented visual speech can disambiguate an acoustically ambiguous plosive between /b/ and /d/ by indicating whether the presented sound was a bilabial or an alveolar sound. Listeners who have been exposed to audiovisual stimuli containing an auditory idiosyncrasy show boundary shifts that are in line with the visual disambiguating information in a subsequent auditory-only categorization task. Auditory phonetic categories are thus retuned both by lexical information and by simultaneously presented visual speech information, the effects of which have also been shown to be statistically similar in size (Van Linden and Vroomen, 2007).

Visual speech itself can also be idiosyncratic, however. Familiarity with the visual speech of a talker can improve subsequent recognition of the talker's visual and auditory speech (Rosenblum *et al*., 2000; Yakel *et al*., 2000; Rosenblum *et al*., 2007). Participants recognized visual speech better, for example, when the same speaker was presented throughout a visual-only recognition task than when multiple speakers were shown (Yakel *et al*., 2000). Listeners can also match a speaker's face producing a sentence to their subsequently presented voice, even when the linguistic content of the visual and auditory speech differ (Kamachi *et al*., 2003; Lander *et al*., 2007). These results suggest that listeners adjust to the visual idiosyncrasies of a speaker. Auditory speech information can guide the adjustment to visual idiosyncrasies when these make visual productions of sounds ambiguous. Baart and Vroomen (2010) presented listeners with videos of a talker producing /o?so/, where /?/ was a visually ambiguous nasal between /m/ and /n/. Audiovisual stimuli were created by combining the ambiguous visual speech input with natural auditory /omso/ or /onso/ tokens. Exposure to these audiovisual stimuli resulted in retuning of the visual phonetic categories. Auditory information thus guides retuning of visual phonetic categories, confirming that speech information from one modality can change category boundaries in the other modality.

However, listeners may also apply lexical knowledge to adjust visual phonetic categories, either by using lexical knowledge to retune visual categories directly, or by applying what they learn about a talker's auditory speech to adjust their expectations about the talker's visual speech. Applying lexical information to audiovisual speech could well be useful for listeners, as idiosyncrasies do not necessarily occur only in one modality at a time. In fact, given the links between visible articulatory movements and the resulting auditory sounds (Yehia *et al*., 1998), idiosyncrasies that are both auditorily and visually expressed are probable. In such cases, with both modalities containing an idiosyncrasy, there would be no opportunity for one modality to guide retuning of phonetic categories in the other. In Experiment 1, we tested whether lexical knowledge can disambiguate audiovisually idiosyncratic speech and whether visual phonetic categories can be retuned on the basis of this lexical knowledge.

We also tested whether the retuning of visual phonetic categories can occur through generalization across the modalities. If auditory and visual phonetic categories are tightly linked, then listeners should be able to retune their visual categories even if no visual information about the idiosyncrasy was present during exposure. The retuning of

auditory phonetic categories would generalize across modalities and therefore indirectly affect visual phonetic categories. Visual-only exposure to the speech of a particular speaker has been shown to facilitate subsequent recognition of that speaker's auditory-only speech, both in a long-term priming task and in a sentence-recognition task (Kim *et al.*, 2004; Rosenblum *et al.*, 2007). Rosenblum and colleagues (2007), for instance, asked listeners to lip-read a speaker for about one hour before being asked to recognize speech in noise. Listeners who heard the same speaker in the recognition task as they had seen during the exposure task performed better than listeners who heard a different speaker in the two tasks. Listeners are thus able to extract speaker-specific information from one modality and apply it to the recognition of speech in another modality. Transfer of speaker-specific knowledge across modalities has not yet been shown for phonetic retuning, however, and it remains unclear whether changes in the auditory phonetic categories could also bring about changes in the visual phonetic categories. (Certainly unambiguous auditory information can guide the retuning of visual categories; Baart and Vroomen, 2010). In Experiment 2, we therefore tested the possibility for lexically guided retuning of auditory phonetic categories to generalize across modalities. Visual category boundaries would then be affected by lexical information, even though the listener had not received visual information about the speaker's idiosyncrasy.

Thus in Experiment 1, two groups completed multiple repetitions of an audiovisual lexical decision task, each directly followed by visual-only categorization. During the lexical decision task, one group heard and saw an ambiguous speech token between /p/ and /t/ that replaced all word-final /p/ tokens. Another group heard and saw the same ambiguous token replacing natural /t/ tokens. In a subsequent categorization task, both groups categorized steps from a visual-only Dutch nonword continuum from /soːp/ to /soːt/. In Experiment 2, exposure was as in Experiment 1, but both groups only heard the exposure speaker. In the categorization test phases, both groups again categorized steps from the visual /p/-/t/ continuum. At the end of Experiment 2, both groups then also categorized steps from an auditory /p/-/t/ continuum. If lexical knowledge (directly or indirectly) retunes visual phonetic categories, then we should observe a shift in the visual phonetic boundaries in Experiment 1. If lexically guided retuning of auditory phonetic categories further generalizes across modalities, a similar shift should be seen in Experiment 2, despite the absence of visual speech information during the lexical decision task. This would mean that lexical knowledge retuned auditory categories, which in turn changed the visual categories.

II. EXPERIMENT 1

A. Method

1. Participants

Forty-two native speakers of Dutch (average age 20.5 yr; six males) were paid for their participation. All participants reported normal hearing and had normal or corrected-to-normal vision. Two participants were excluded due to their insensitivity to the auditory-only continuum in the pretest. Another ten participants (four in the /p/-exposure group and six in the /t/-exposure group) were excluded for failing to exceed a threshold of 50% correct "word" responses to the ambiguous target words on the lexical decision task. The final data set that was analyzed consisted of data from 30 participants, 16 in the /p/-exposure group and 14 in the /t/-exposure group. Fifteen additional participants from the same population took part in a visual-only pilot experiment.

2. Materials

Four /p/-final (*hoop*, *kroop*, *zoop*, and *siroop*) and four /t/-final Dutch words (*groot*, *schoot*, *schroot*, and *vergroot*) were selected as target words for the exposure phase. None of these eight target words formed a word when its coda was replaced with any other phoneme from the same viseme category (Van Son *et al.*, 1994; e.g., *hoop* is a Dutch word, but *hoot*, *hoob*, and *hoom* are not) or with the respective other plosive. Target words contained no other phonemes from the relevant viseme categories and no other instances of /p/ or /t/. In both word sets, one target word was disyllabic and the other three were monosyllabic. Word sets were matched on their mean frequency, number of syllables, and their lexical stress patterns using the CELEX lexical database (Baayen *et al.*, 1993). Eight phonotactically legal nonsense words were created that ended in either /f/ or /x/. These eight nonsense words contained no phonemes from the viseme categories of the target plosives. In all 16 items (8 target words and 8 nonsense words) the same vowel, /oː/, preceded the final phoneme. For the categorization tasks, the nonsense words /soːp/ and /soːt/ were used.

A male native speaker of Dutch was video recorded with a Sony (Sony Corporation, Tokyo, Japan) DCR-HC1000E camera. Audio was recorded with two standalone Sennheiser (Sennheiser electronic GmbH & Co. KG, Hanover, Germany) microphones. Videos showed the speaker's head and the top of his shoulders. The speaker produced the target words both with their natural word-final plosive and with the alternative plosive (e.g., the Dutch word *kroop* and its nonsense word counterpart *kroot*). The same speaker also produced the eight nonsense words for the lexical decision task and the *soop* and *soot* items for the categorization tasks. All items were recorded in pairs and the talker was instructed to avoid list intonation. Videos were digitized as uncompressed 720×576 .avi (audio video interleave) files in PAL format. Audio sampling rate was 44.1 kHz.

We created an auditory-only continuum and a visual-only continuum using the same audiovisual *soop* and *soot* tokens for both continua. The visual-only continuum was created for the visual-only pretest and posttests. The auditory-only continuum was presented in the auditory-only pretest that was conducted to find each individual participant's most ambiguous auditory step ($A_?$). The selected sound $A_?$ appeared in all ambiguous target words for that participant during exposure. It was presented together with a visually ambiguous final plosive $V_?$ in these words. The ambiguous visual token was the same across participants but different for each target word.

a. Auditory-only pretest materials. An audiovisual token of each of *soop* and *soot* was selected based on how well the two tokens could be merged visually without causing any noticeable blurring of the speaker's facial features and facial contour. The auditory signal from both tokens was extracted and edited using Praat (Boersma, 2001). The word-final plosives were excised by removing all sound up to the first zero crossing of the release burst. The releases of the two plosives were then morphed using the STRAIGHT signal-processing package (Kawahara *et al.*, 1999) for Matlab (The MathWorks, Inc., Natick, MA). This resulted in 21 individual plosive releases changing in equal 5% steps from an unambiguous auditory /t/ release (0% /p/) to an unambiguous auditory /p/ release (100% /p/). In order to provide an unbiased context for the edited releases, an ambiguous *soo* token was created by removing the closure duration and the release from the auditory *soop* and *soot* tokens. The two resulting *soo* tokens were then morphed in a seven-step continuum with STRAIGHT. The middle step (step 4) was selected as the ambiguous context and was then combined with all 21 morphed releases. Since neither the ambiguous context nor the morphed releases contained a closure duration, a stretch of complete silence was added to these continuum steps in Praat. This artificial closure duration was manipulated to be the same duration as the average duration of the closure for /p/ and /t/ in the original *soop* and *soot* tokens (1652 ms and 1542 ms, respectively; 1588 ms for the continuum steps).

b. Visual-only pretest and posttest materials. The audiovisual tokens that were used to create the visual-only *soop-soot* continuum were the same as for the auditory-only continuum. To create the visual-only continuum, the video tracks of the *soop* and *soot* tokens were edited using Adobe (Adobe Systems, Mountain View, CA) Premiere CS3. These video tracks were overlaid and the opacity level of the /p/ video was systematically varied. A clip with 0% opacity for the /p/-final video shows the speaker producing an unambiguous /t/, while a clip with 100% opacity for the /p/-final videos shows the speaker producing an unambiguous /p/. A 21-step visual-only continuum was created that ranged from 0% opacity for /p/ (i.e., an unambiguous /t/ token) to 100% opacity for /p/ (i.e., an unambiguous /p/ token) by increasing the opacity for /p/ in increments of 5%.

c. Audiovisual exposure materials. Audiovisual exposure items consisted of eight natural target words ending in /p/ or /t/ and eight natural nonsense words ending in /f/ or /x/. In addition, eight ambiguous versions of these target words were created with auditorily and visually ambiguous final plosives. To create the visually ambiguous plosives, we selected two audiovisual tokens for each target word (i.e., the target word and the same word ending in the alternative plosive) on the basis of how well they could be merged visually. For each of the eight target words, a visual-only and auditory-only continuum was created using the same stimulus creation procedures detailed for the auditory and visual pretest materials. The most ambiguous visual step for each target word ($V_?$) was established on the basis of a pilot study and was the same across participants, but different across target words. The video containing this step was combined with an audio track containing each participant's most ambiguous auditory step ($A_?$), as found in the auditory-only pretest for each participant. This created target words in which the critical sounds were ambiguous in both modalities ($A_?V_?$).

d. Visual-only pilot. A pilot study was conducted to test participants' sensitivity to the visual-only *soop-soot* continuum and to select the most ambiguous visual continuum step for each of the eight target words. Participants categorized 13 steps from the *soop-soot* continuum (steps 0, 15, 30, 35, 40, 45, 50, 55, 60, 65, 70, 85, 100). Participants also categorized ten steps (steps 0, 15, 30, 35, 40, 45, 50, 55, 60, 65, 70, 85, 100) from four of the eight target word continua. The four target word continua always consisted of two /p/-final targets and two /t/-final targets, assigned randomly to each participant. The *soop-soot* continuum was always presented first. The presentation order of the following four target-word continua was rotated across lists. For every continuum, each step was repeated eight times in a newly randomized order within each repetition. The two response alternatives (i.e., /p/ or /t/) were displayed on a computer screen beneath the video of the speaker producing an utterance. Stimuli were presented 200 ms after trial onset. Participants were instructed to respond as accurately and as quickly as possible by pressing one of the two buttons on a button box that corresponded with the "p" and "t" labels shown on the computer screen. Each new trial started only after participants had given a response. No feedback was provided.

The results of the pilot study can be seen in Figs. 1(a) and 1(b). Figure 1(a) shows the results for the visual-only *soop-soot* continuum and Fig. 1(b) shows the results for the visual-only target-word continua. The results indicate that participants were sensitive to the visual-only continua for both *soop-soot* and the target words and gave more [p] responses the more /p/-like the continuum step. The most ambiguous visual continuum step for each of the eight target words was selected on the basis of the 50% cut-off points, indicated by the vertical lines in Fig. 1(b). These steps were chosen as $V_?$ for the creation of the audiovisual exposure versions of these target words. Whenever the 50% point fell between two categorized steps, a new video was created with a step that was between the two steps adjacent to the 50% point. Four of the target stimuli contained such a newly created step (*kroop*, *zoop*, *goot*, and *schoot*). The selected steps for these target words were 52, 54, 51, and 43, respectively [cf. Fig. 1(b)].

3. Design and procedure

Participants were randomly assigned to either the /p/-exposure group or the /t/-exposure group and tested individually in a sound-attenuated booth. The experimental session lasted 45 min. Participants started the experiment with an auditory-only pretest in which they categorized 15 steps from the auditory-only *soop-soot* continuum (steps 1, 4, 6–16, 18, 21). All continuum steps were presented eight

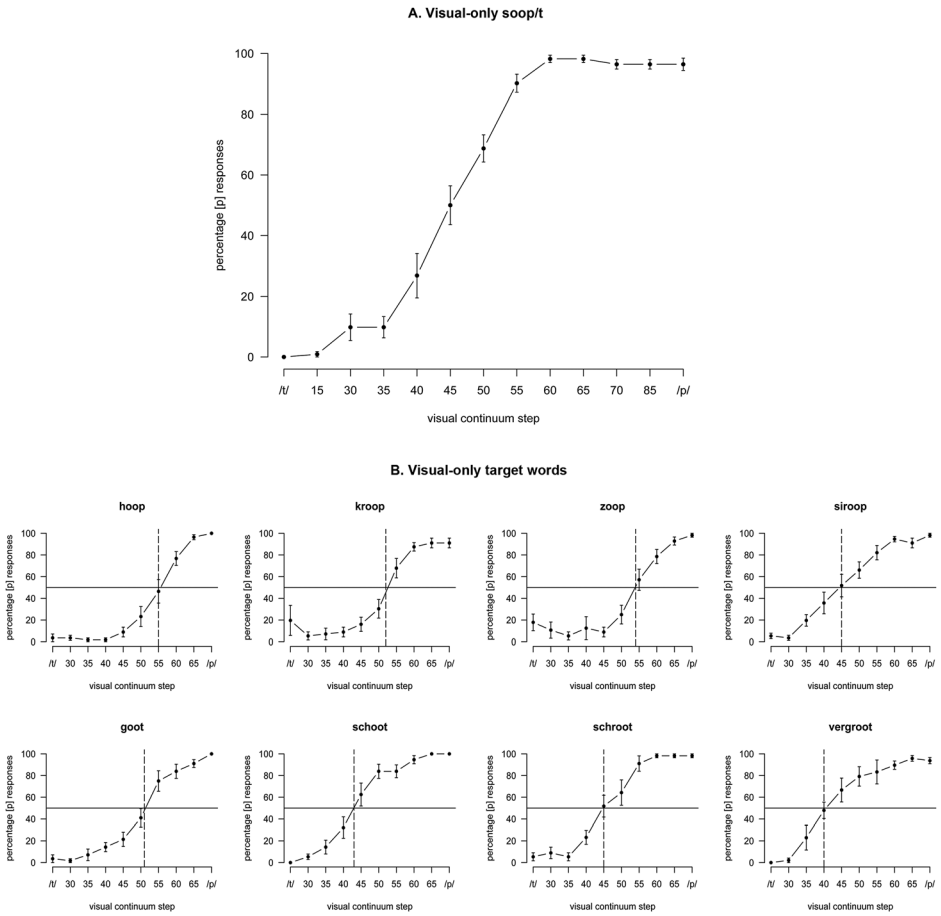

FIG. 1. Mean percentages of [p] responses as a function of /so:p/-/so:t/ continuum steps (Panel A) and for the visual-only continua of all eight target words (Panel B) in the visual-only pilot study. Horizontal lines mark 50% [p] responses. Vertical lines mark the visual step used to create the audiovisual exposure materials. Error bars show the standard error of the mean.

times in a newly randomized order for every repetition. The audio was presented over Sennheiser HD280 headphones at a fixed level. Participants indicated whether the final sound they had heard was /p/ or /t/ by clicking with the computer mouse on labeled buttons on a computer screen. Each new trial started 500 ms after a response had been given. The results for the auditory-only pretest were used to select each participant's most ambiguous auditory token $A_?$ for use in the rest of the experiment. $A_?$ was always the step closest to participants' 50% cut-off point between [p] and [t].

After the auditory-only pretest, participants performed a visual-only pretest. Participants categorized seven steps from the visual-only *soop-soot* continuum (steps 0, 35, 40, 45, 50, 55, 100). Each step was presented three times with presentation blocked by repetition. Participants indicated whether the final sound the talker had produced was a /p/ or a /t/ by pressing the button on a button box that corresponded to the respective labels shown on-screen. New trials started 800 ms after participants gave a response. This visual-only pretest provided a baseline to which the posttest results were compared.

The exposure phase consisted of an audiovisual lexical decision task. Each exposure block was immediately followed by another visual-only categorization block (posttest) and participants completed a total of ten repetitions of such exposure-posttest sequences. Participants received four /t/-final and four /p/-final target words, intermixed with four /f/-final and four /x/-final nonsense words in each exposure block. Participants assigned to the /p/-exposure group received /p/-final target words where the final plosive was both visually and auditorily ambiguous ($A_?V_?$) along with natural /t/-final target words (A_tV_t). Participants in the /t/-exposure group received auditorily and visually ambiguous /t/-final words ($A_?V_?$) along with natural /p/-final words (A_pV_p). The exposure condition was the same for a participant across all repetitions of the exposure and posttest phases. $A_?$ in the audiovisual exposure materials was selected on the basis of each participant's pretest results and the same in all words. $V_?$ in the materials was selected based on the pilot study data and the same for all participants in a given word, but different across words. Participants watched and heard the speaker produce each item and indicated as quickly and as accurately as possible whether or not what the talker had said was an existing Dutch word. Answers were provided by pressing the button on a button box that corresponded with the respective label shown on the computer screen ("w" for "wel"/"yes"; "n" for "niet"/"no"). All 16 items were presented twice in random order blocked by repetition. New trials started 800 ms after the participant gave a response.

B. Results and discussion

Results were analyzed using linear mixed-effect models in the R statistical program (Version 2.11.0; R Development Core Team, 2007) by using the lmer function of the lme4 library (Bates and Sarkar, 2007). The dependent variable for the exposure phase was the binomial word judgment (correct or incorrect). The dependent variables for the pretest and

posttests were the binomial response to the continuum steps (0 = /t/; 1 = /p/). A logistic linking function was used for these categorical dependent variables. The best-fitting model for each data set was established through systematic model comparison using likelihood-ratio tests. We always started with the full model, gradually removing factors that did not contribute to a better model fit, starting with the factors with the largest p values. Main effects were only removed if their factors did not contribute to an interaction. All best-fitting models included participants as a random factor. Group (/p/-exposure group vs /t/-exposure group) was evaluated as a contrast-coded fixed factor in all analyses. Ambiguity (natural target words vs ambiguous target words) was evaluated as a contrast-coded fixed factor in the analysis of the exposure data. Visual continuum step was evaluated as a numerical factor centered on the middle step in the pretest and the posttest analyses. Test (pretest vs posttest) was evaluated as a contrast-coded fixed factor in the comparison of the visual-only pretest and posttest data.

1. Visual-only pretest

There was no difference in the number of [p] responses given by the two groups at pretest (not a predictor, $\beta = -0.31$, standard error (SE) = 0.48, $p = 0.52$). Both groups gave more [p] responses to the more /p/-like visual tokens ($\beta = 0.20$, SE = 0.01, $p < 0.001$; see Fig. 2). This indicates that the two groups were sensitive to the visual-only continuum and did not differ prior to testing in their visual categories.

2. Audiovisual exposure

Table I (upper row) gives the mean percentages of correct "word" responses to ambiguous and nonambiguous versions of the target words. Participants gave more correct responses to the natural target words than to the target words containing an ambiguous plosive ($\beta = 0.77$, SE = 0.13, $p < 0.001$). This difference between natural and ambiguous target words was numerically larger in the /p/-exposure group (natural: 94%, ambiguous: 88%) than in the /t/-exposure group (natural: 95%, ambiguous: 93%), but the interaction was only marginally significant [$\chi^2(1) = 3.58$, $p = 0.058$].

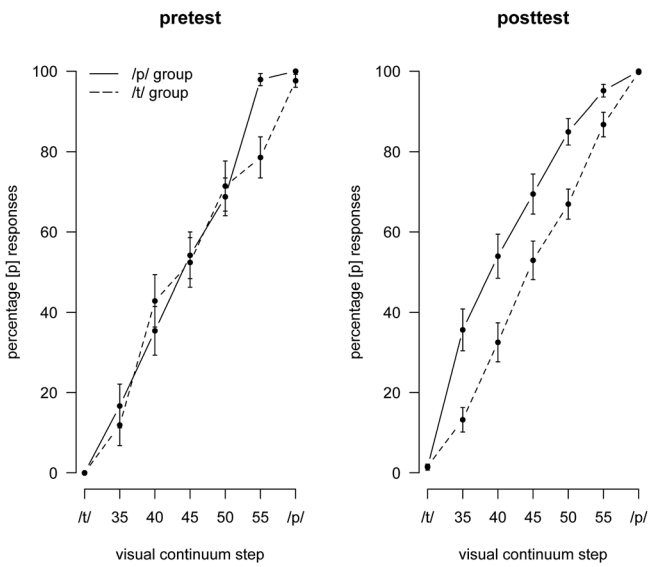

FIG. 2. Mean percentages of [p] responses across pretest and posttests as a function of visual continuum step in Experiment 1. Solid lines show the results for the /p/-exposure group and dashed lines the results for the /t/-exposure group. Error bars show the standard error of the mean.

TABLE I. Mean percent correct responses to natural and ambiguous /p/-final and /t/-final words in Experiments 1 and 2.

	Natural		Ambiguous	
	/p/ words	/t/ words	/p/ words	/t/ words
Experiment 1	95.44	94.44	87.50	93.06
Experiment 2	92.45	96.30	81.76	95.31

3. Visual-only posttests

The data from all visual-only posttest blocks were pooled together since there was no effect of block ($\beta = -0.00$, SE = 0.01, $p = 0.96$). Participants gave more [p] responses to the more /p/-like visual continuum steps in the posttest, again indicating sensitivity to the visual-only continuum ($\beta = 0.19$, SE = 0.01, $p < 0.001$). Participants in the /p/-exposure group gave more [p] responses than participants in the /t/-exposure group ($\beta = -1.21$, SE = 0.50, $p < 0.05$). This result indicates an effect of learning in line with exposure. Lexical knowledge can thus be used to retune visual phonetic categories. Participants in the /p/-exposure group gave more [p] responses in the posttests than in the pretest ($\beta = 0.92$, SE = 0.19, $p < 0.001$). The responses from participants in the /t/-exposure group in the posttests did not differ from the pretest [$\chi^2(1) = 1.49$, $p = 0.22$]. This indicates that, while there is a difference between the two groups in line with their exposure, this difference between the groups is mainly due to learning in the /p/-exposure group.

III. EXPERIMENT 2

In Experiment 1, we showed that lexical knowledge can be used to shift the boundaries of visual phonetic categories. Exposure to an audiovisually ambiguous sound within a biasing lexical context resulted in a shift of the visual category boundary. This shift was only observed for the /p/-exposure group, but not for the /t/-exposure group. Listeners in Experiment 1 could either have used lexical knowledge to retune visual phonetic categories directly, or used lexical information to retune auditory category boundaries, which in turn influenced visual category boundaries. The observed shift for the visual category boundaries could in the latter case reveal generalization across modalities. In Experiment 2, we directly tested whether retuning of the visual phonetic categories can occur through generalization of speaker knowledge across modalities. In Experiment 2, participants were exposed to auditory-only versions of the audiovisual stimuli of Experiment 1 and were subsequently tested on the visual-only continuum and on an auditory-only version of that continuum. This way, we investigated whether retuning of visual categories can still occur even when visual speech was not presented with the lexically disambiguating context.

A. Method

1. Participants

Forty-four new participants (average age 20.8 yr; 12 males) from the same population as for Experiment 1 were tested. Five participants were excluded due to insensitivity to the auditory continuum during the pretest. An additional eight participants were excluded for failing to exceed a threshold of 50% correct "word" responses to the ambiguous target words on the lexical decision task. All of these excluded participants had been assigned to the /p/-exposure group. The final data set consisted of data from 31 participants, 15 in the /p/-exposure group and 16 in the /t/-exposure group.

2. Materials

Materials for Experiment 2 were the same as those used in Experiment 1. However, rather than audiovisual stimuli, participants received auditory-only versions of the stimuli during the exposure phase. The auditory-only stimuli were created by blacking out the video of the audiovisual stimuli used during exposure in Experiment 1. Stimuli were otherwise identical. The auditory-only posttest stimuli were a subset of the steps of the auditory-only /soːp/-/soːt/ continuum used in the pretest.

3. Design and procedure

There were two differences between the procedure of Experiments 1 and 2. In Experiment 2, the exposure materials were auditory-only rather than audiovisual, and participants performed an additional auditory-only posttest at the end of the experiment. Otherwise, the procedure of Experiment 2 was the same as in Experiment 1. First, an auditory-only pretest established each participant's most ambiguous auditory step ($A_?$) for exposure. Participants then completed ten exposure-posttest repetitions where they first performed an auditory-only lexical decision task (exposure) and then a visual-only categorization task (posttest). After these exposure-posttest repetitions, participants completed an additional auditory-only categorization task. This auditory test was added as a control to test whether the exposure materials would lead to retuning of auditory phonetic categories. It was conducted at the end of testing to ensure comparability between the visual-only posttest results for Experiments 1 and 2.

The auditory-only posttest consisted of three steps from the auditory-only *soop-soot* continuum, namely the participant's most ambiguous step, $A_?$, and a more /p/-like step, $A_{?-1}$, and a more /t/-like step, $A_{?+1}$. All three steps were presented eight times in a newly randomized order for each repetition. Participants responded by pressing one of the buttons on a button box that corresponded to the labels shown on the computer screen.

B. Results and discussion

Results were analyzed as for Experiment 1. Group (/p/-exposure group vs /t/-exposure group) was evaluated as a contrast-coded fixed factor and auditory continuum step as a fixed factor centered on the middle step in the analysis of the auditory-only posttest data. Participants were included as a random factor in the best-fitting model for the auditory-only posttest.

1. Visual-only pretest

The two groups did not differ in the number of [p] responses given in the visual-only continuum steps at pretest [not a predictor, $\chi^2(1) = 0.10$, $p = 0.75$]. Both groups were sensitive to the visual-only continuum and gave more [p] responses the more /p/-like the visual continuum steps were ($\beta = 0.18$, SE = 0.02, $p < 0.001$). This indicates that the two groups were sensitive to the visual-only continuum but their visual categories did not differ prior to exposure.

2. Auditory-only exposure

There was no difference between the responses of the /p/-exposure group and the /t/-exposure group in the exposure phase (not a predictor, $\beta = 0.44$, SE = 0.52, $p = 0.39$; see Table I, lower row). Overall, participants gave more correct responses to the natural target words than to the ambiguous target words ($\beta = 0.77$, SE = 0.13, $p < 0.001$). The difference between responses to the natural and ambiguous target words for the /p/-exposure group (natural: 96%; ambiguous: 82%) was opposite to that observed for the /t/-exposure group (natural: 92%: ambiguous: 95%; $\beta = -2.55$, SE = 0.27, $p < 0.001$). The /p/-exposure group gave more correct responses to the natural target words than to the ambiguous target words ($\beta = 1.98$, SE = 0.19, $p < 0.001$), while the /t/-exposure group gave fewer correct responses to the natural target words than to the ambiguous target words ($\beta = -0.57$, SE = 0.19, $p < 0.01$). The unexpected pattern for the /t/-exposure group may have been due to the unambiguous item *zoop*, which had been rejected as a word in 42% of all presentations. This item may have been

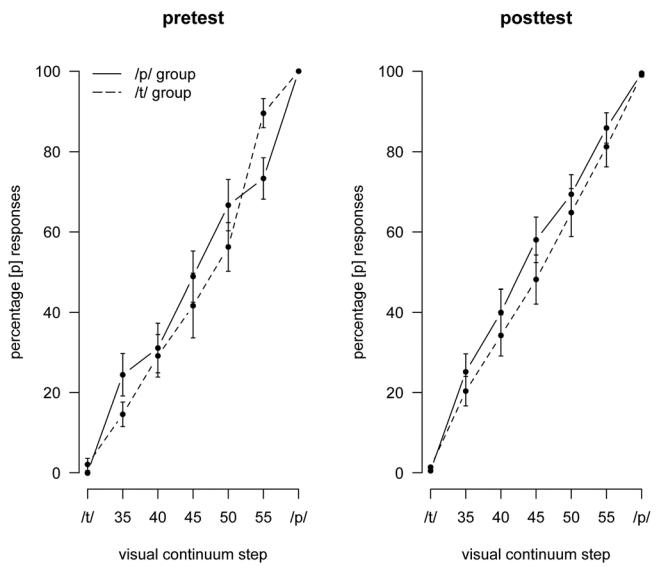

FIG. 3. Mean percentages of [p] responses across pretest and posttests as a function of visual continuum step in Experiment 2. Solid lines show the results for the /p/-exposure group and dashed lines the results for the /t/-exposure group. Error bars show standard error of the mean.

categorized as a nonword, since participants may have thought of it as being too colloquial or dialectal to be a real Dutch word.

3. Visual-only posttests

The results from the visual-only posttest revealed no differences between the number of [p] responses given by the two groups [not a predictor, $\chi^2(1) = 0.51$, $p = 0.47$], indicating that auditory-only exposure did not affect the subsequent categorization of the visual-only continuum (see Fig. 3). Participants thus did not retune their visual category boundaries after auditory-only exposure. Participants in both groups were sensitive to the visual-only continuum and gave more [p] responses the more /p/-like the continuum step was ($\beta = 0.18$, SE = 0.01, $p < 0.001$).

4. Auditory-only posttest

Overall, participants were sensitive to the auditory-only continuum and gave more [p] responses to the more /p/-like steps ($\beta = 0.50$, SE = 0.11, $p < 0.001$; see Fig. 4). Participants in the /p/-exposure group gave more [p] responses than those in the /t/-exposure group ($\beta = -1.38$, SE = 0.56, $p < 0.05$), indicating that the categorization of the auditory-only posttest was influenced by exposure. This finding replicates results reported by earlier studies by showing that lexical information can guide retuning of auditory phonetic categories (Norris *et al.*, 2003; McQueen *et al.*, 2006a; McQueen *et al.*, 2006b). Taken together, the results of the auditory-only posttest and the visual-only posttests show that while listeners used lexical information here to retune their auditory phonetic categories based on the auditory-only exposure, this retuning did not affect visual phonetic categories.

IV. GENERAL DISCUSSION

Listeners perceive speech bimodally when they hear and see someone talk. Idiosyncrasies of a speaker expressed in one modality can be disambiguated by information in the simultaneously presented speech in the other modality (Bertelson *et al.*, 2003; Baart and Vroomen, 2010). This disambiguation leads to the retuning of category boundaries in line with the disambiguating context. Listeners also use their lexical knowledge to retune auditory phonetic categories to talker idiosyncrasies contained in auditory speech (Norris *et al.*, 2003). The results of the present study show that lexical knowledge can also retune visual phonetic categories. Exposure to audiovisually ambiguous sounds that were disambiguated by lexical information resulted in shifts of listeners' visual category boundaries. Furthermore, the current results also indicate that visual phonetic categories are only influenced by lexical knowledge when visual information about the idiosyncrasy was available to the listener. Auditory-only exposure to an idiosyncratic sound resulted in retuning of auditory phonetic categories, but did not affect visual phonetic categories. Phonetic retuning in one modality does not generalize to the categories in another modality.

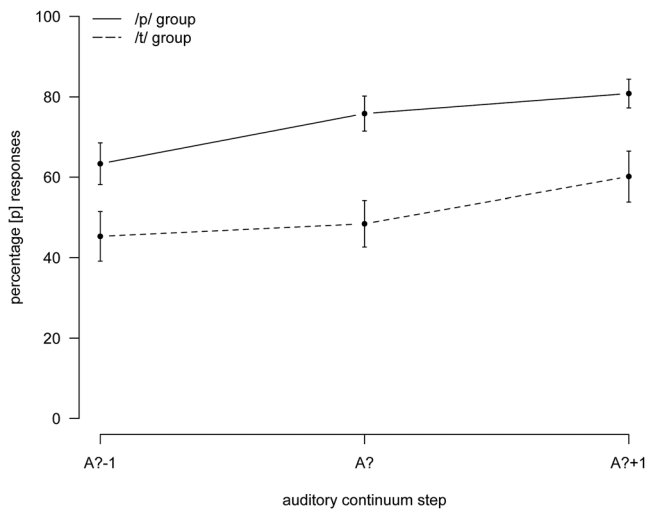

FIG. 4. Mean percentages of [p] responses for the auditory-only posttest in Experiment 2 as a function of auditory continuum step. Solid lines show the results for the /p/-exposure group and dashed lines the results for the /t/-exposure group. Error bars show the standard error of the mean.

Listeners use their lexical knowledge to adjust visual category boundaries to optimize speech recognition. Retuning the visual phonetic categories in this way is particularly beneficial in situations where the same idiosyncrasy is observed in both the auditory and the visual modality. In such cases, information from neither modality can be used to guide the perceptual learning. Listeners are then dependent on other sources, such as their linguistic knowledge, for the resolution of the ambiguity in the audiovisual speech input. Listeners use lexical knowledge to directly retune their visual phonetic categories, or do so indirectly via the retuning of auditory categories. Our results show, however, that listeners were only able to adjust their visual category boundaries if the lexicon disambiguated the visual idiosyncrasy as /p/. This could indicate that retuning of the category boundaries only occurs for those phonemes that are strongly defined visually (here, the bilabial plosives), but not for those phonemes that are difficult to identify visually (here, the alveolars, for which the defining place of articulation is inside the oral cavity). That is, a departure from typicality that is not readily noticeable to the eye will not prompt category retuning. Although only further research will conclusively decide the issue, listeners may only be sensitive to speaker idiosyncrasies in phonemes that are visually distinct, and in consequence, it may be only such phoneme categories that are retuned.

It should be noted that the ability to resolve visual ambiguity by reference to existing knowledge, and apply learning from such ambiguity resolution to future visual perceptual processing, is by no means confined to speech recognition. The interpretation of color in visual processing involves similar perceptual learning operations as a color-perception analog of the Norris *et al.* (2003) experiment showed. Mitterer and de Ruiter (2008) for example presented viewers with pictures of fruit, typically encountered either in yellow or orange, in an ambiguous color between yellow and orange, and then collected categorization judgments on a yellow-orange continuum of colored socks. Viewers who had seen the ambiguous color on bananas judged more socks along the continuum as yellow, whereas viewers who had seen the ambiguous color on oranges categorized more socks as

orange. The same kind of visual category shift was also observed with ambiguous letters between H and N presented word-finally in sequences such as WEIG- versus REIG- (Norris *et al.*, 2006). In our complex world, sensory processing in any modality is liable to deliver ambiguous input, but our cognitive processing is able to resolve the ambiguity by referring to knowledge of many sorts, and can learn from this to improve future processing.

The results of Experiment 2 provide evidence that the visual phonetic categories were only influenced by listeners' lexical knowledge if visual information about the speaker's idiosyncrasy was available to the listener. Phonetic retuning occurred for listeners' auditory phonetic categories after exposure to auditory-only idiosyncratic speech, but no such retuning was observed for the visual phonetic categories. Lexically guided retuning in one modality thus did not generalize to another modality and the boundary shifts for the visual phonetic categories in Experiment 1 must have occurred because listeners obtained information about how to retune their visual categories directly from seeing the speaker talk. For retuning to occur, information about the idiosyncrasy needs to be available to the listener from the modality for which the phonetic categories are retuned.

Transfer for speaker information across modalities has been observed in a previous study, however (Rosenblum *et al.*, 2007). Rosenblum and colleagues found transfer of knowledge about a speaker's visual speech to their auditory speech. A variety of methodological differences between the Rosenblum study and the current study could provide an explanation for the discrepancy in the findings. Most notably, participants in the Rosenblum study received the critical words in sentences during exposure and test. In our study, participants were presented with isolated words during exposure and nonsense syllables during test. Words are generally more easily identified when presented in a meaningful sentence context than when they are presented in isolation (Miller *et al.*, 1951; Boothroyd and Nittrouer, 1988; Grant and Seitz, 2000). This, in addition to the increased amount of exposure in the Rosenblum study compared to our study, could have led to better learning and therefore cross-modal transfer of speaker information. But because words were presented in sentences, listeners in the Rosenblum study could also arguably have been familiarized with, and subsequently have generalized, different properties of the speaker than listeners in our study. Speaker familiarity established on the basis of sentences does not significantly improve subsequent recognition of novel words in isolation (Nygaard and Pisoni, 1998), indicating that listeners may tune in to a different set of speaker-specific properties depending on the exposure materials. Sentences provide information about speaker-specific properties such as prosody, duration, and speaking rate (Grant *et al.*, 1998; Nygaard and Pisoni, 1998; Adank and Janse, 2009), to which listeners can attune, but which are not available from isolated words. Learning of these speaker characteristics could possibly transfer across modalities (see, for instance, Cvejic *et al.*, 2012), while learning of phonetic idiosyncrasies, as tested in our study, may not.

Retuning for auditory and visual phonetic categories thus appears to reflect two distinct processes that do not necessarily affect one another. Listeners retune their boundaries for whichever category is problematic during exposure to a speaker, considering all available information. If speech from only one modality is provided, then only the boundaries of categories for that modality are changed and this shift does not affect the category in the other modality. Retuning for the visual category failed in Experiment 2 because the ambiguity was only presented in the auditory modality and so listeners were not aware of how to retune their visual category boundary. This finding indicates that auditory and visual categories are not inextricably linked and that changes for the categories in one modality do not necessarily result in changes for the categories in the other modality.

The results of Experiment 2 pose a potential problem for theories that posit that listeners use information about the speaker's intended vocal tract gestures for speech perception, i.e., motor theory and direct realist theory (Liberman *et al.*, 1967; Liberman and Mattingly, 1985; Fowler, 1986,1991; Fowler *et al.*, 2003; Galantucci *et al.*, 2006). In these theories, it is postulated that listeners are able to obtain information about the underlying gestures from auditory speech input. If this were the case, then listeners should be able to retune their visual phonetic categories based on auditory speech alone. That is, if lexical knowledge disambiguates an auditory speaker idiosyncrasy, then the auditory speech signal alone should contain all the information necessary to retune the characteristic articulatory features that encompass the corresponding visual phonetic category. The finding that lexically guided retuning of auditory categories does not transfer to visual categories in Experiment 2 suggests, however, that such information about the articulatory movements is not extracted (or directly perceived) from the auditory speech input. Instead, auditory-only presentation results in boundary shifts only for auditory phonetic categories.

In the present experiments, we have shown that reference to information outside the speech signal itself is deployed for visual as for auditory ambiguity resolution. Such information can be lexical, as in the present experiments and in many others, but it need not be; for instance, phonotactic constraints realized in nonword sequences also lead to similar learning (Cutler *et al.*, 2008). Our study indicates that while there is a tight link between auditory and visual speech, the respective categories are separate and retuning of each is a separate process.

V. CONCLUSIONS

The present study extends our knowledge about lexically guided retuning of phonetic categories. First, we have demonstrated that lexical information can guide retuning of visual phonetic categories. Second, lexical information does not retune visual categories through generalization across modalities. Despite the inherent link between auditory and visual speech, listeners do not adjust their visual category boundaries on the basis of lexically retuned auditory category boundaries. Retuning based on lexical information helps learning about the idiosyncrasies in the modality they occur in, but does not generalize across modalities.

Adank, P., and Janse, E. (**2009**). "Perceptual learning of time-compressed and natural fast speech," J. Acoust. Soc. Am. **126**, 2649–2659.
Baart, M., and Vroomen, J. (**2010**). "Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading," Neurosci. Lett. **471**, 100–103.
Baayen, R. H., Piepenbrock, R., and Van Rijn, H. (**1993**). *The Celex Lexical Database* [cd-rom] (Linguistic Data Consortium, University of Pennsylvania, Philadelphia).
Bates, D., and Sarkar, D. (**2007**). "lme4: Linear mixed-effect models using S4 classes," http://lme4.r-forge.r-project.org (Last viewed 03/07/2013).
Bertelson, P., Vroomen, J., and de Gelder, B. (**2003**). "Visual recalibration of auditory speech identification: A McGurk aftereffect," Psychol. Sci. **14**, 592–597.
Boersma, P. (**2001**). "Praat, a system for doing phonetics by computer," Glot Int. **5**, 341–345.
Boothroyd, A., and Nittrouer, S. (**1988**). "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. **84**, 101–114.
Cutler, A., McQueen, J. M., Butterfield, S., and Norris, D. (**2008**). "Prelexically-driven perceptual retuning of phoneme boundaries," in *Proceedings of Interspeech: 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, p. 2056.
Cvejic, E., Kim, J., and Davis, C. (**2012**). "Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody," Cognition **122**, 442–453.
Eisner, F., and McQueen, J. M. (**2005**). "The specificity of perceptual learning in speech processing," Percept. Psychophys. **67**, 224–238.
Eisner, F., and McQueen, J. M. (**2006**). "Perceptual learning in speech: Stability over time," J. Acoust. Soc. Am. **119**, 1950–1953.
Foulkes, P., and Docherty, G. (**2006**). "The social life of phonetics and phonology," J. Phonetics **34**, 409–438.
Fowler, C. A. (**1986**). "An event approach to the study of speech perception from a direct realist perspective," J. Phonetics **14**, 3–28.
Fowler, C. A. (**1991**). "Listeners do hear sounds, not tongues," J. Acoust. Soc. Am. **99**, 1730–1741.
Fowler, C. A., Brown, J. M., Sabadini, L., and Welhing, J. (**2003**). "Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks," J. Mem. Lang. **49**, 396–413.
Galantucci, B., Fowler, C. A., and Turvey, M. T. (**2006**). "The motor theory of speech perception reviewed," Psychon. Bull. Rev. **13**, 361–377.
Grant, K. W., and Seitz, P. F. (**2000**). "The recognition of isolated words and words in sentences: Individual variability in the use of sentence context," J. Acoust. Soc. Am. **107**, 1000–1011.
Grant, K. W., Walden, B. E., and Seitz, P. F. (**1998**). "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration," J. Acoust. Soc. Am. **103**, 2677–2690.
Helfer, K. S., and Freyman, R. L. (**2005**). "The role of visual speech cues in reducing energetic and informational masking," J. Acoust. Soc. Am. **117**, 842–849.
Jesse, A., and Massaro, D. W. (**2010**). "The temporal distribution of information in audiovisual spoken-word identification," Atten. Percept. Psycho. **72**, 209–225.
Jesse, A., and McQueen, J. M. (**2011**). "Positional effects in the lexical retuning of speech perception," Psychon. Bull. Rev. **18**, 943–950.
Jesse, A., Vrignaud, N., Cohen, M. M., and Massaro, D. W. (**2000/2001**). "The processing of information from multiple sources in simultaneous interpreting," Interpreting **5**, 95–115.
Kamachi, M., Hill, H., Lander, K., and Vatikiotis-Bateson, E. (**2003**). "'Putting the face to the voice': Matching identity across modality," Curr. Biol. **13**, 1709–1714.
Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based *f*0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.
Kim, J., Davis, C., and Krins, P. (**2004**). "Amodal processing of visual speech as revealed by priming," Cognition **93**, B39–B47.
Kraljic, T., and Samuel, A. G. (**2006**). "Generalization in perceptual learning for speech," Psychon. Bull. Rev. **13**, 262–268.
Lander, K., Hill, H., Kamachi, M., and Vatikiotis-Bateson, E. (**2007**). "It's not what you say but the way you say it: Matching faces and voices," J. Exp. Psychol. Hum. Percept. Perform. **33**, 905–914.
Laver, J., and Trudgill, P. (**1979**). "Phonetic and linguistic markers in speech," in *Social Markers in Speech*, edited by K. R. Scherer and H. Giles (Cambridge University Press, Cambridge), pp. 1–32.
Liberman, A. M., and Mattingly, I. G. (**1985**). "The motor theory of speech perception revised," Cognition **21**, 1–36.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). "Perception of speech code," Psychol. Rev. **74**, 431–461.
Macleod, A., and Summerfield, Q. (**1987**). "Quantifying the contribution of vision to speech perception in noise," Br. J. Audiol. **21**, 131–142.
Massaro, D. W. (**1987**). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry* (Lawrence Erlbaum, Hillsdale, NJ), pp. 1–327.
Massaro, D. W. (**1998**). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle* (MIT Press, Cambridge, MA), pp. 1–474.
Massaro, D. W., and Cohen, M. M. (**1983**). "Evaluation and integration of visual and auditory information in speech perception," J. Exp. Psychol. Hum. Percept. Perform. **9**, 753–771.
Massaro, D. W., and Friedman, D. (**1990**). "Models of integration given multiple sources of information," Psychol. Rev. **97**, 225–252.
McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.
McQueen, J. M., Cutler, A., and Norris, D. (**2006a**). "Phonological abstraction in the mental lexicon," Cogn. Sci. **30**, 1113–1126.
McQueen, J. M., Norris, D., and Cutler, A. (**2006b**). "The dynamic nature of speech perception," Lang. Speech **49**, 101–112.
Miller, G. A., Heise, G. A., and Lichten, W. (**1951**). "The intelligibility of speech as a function of the context of the test materials," J. Exp. Psychol. **41**, 329–335.
Mitterer, H., and de Ruiter, J. P. (**2008**). "Recalibrating color categories using world knowledge," Psychol. Sci. **19**, 629–634.
Mitterer, H., Chen, Y., and Zhou, X. (**2011**). "Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm," Cogn. Sci. **35**, 184–197.
Norris, D., Butterfield, S., McQueen, J. M., and Cutler, A. (**2006**). "Lexically guided retuning of letter perception," Q. J. Exp. Psychol. **59**, 1505–1515.
Norris, D., McQueen, J. M., and Cutler, A. (**2003**). "Perceptual learning in speech," Cogn. Psychol. **47**, 204–238.
Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**, 355–376.
R Development Core Team. (**2007**). "R: A language and environment for statistical computing," http://cran.r-project.org (Last viewed 05/21/2010).
Reisberg, D., McLean, J., and Goldfiend, A. (**1987**). "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lip-Reading*, edited by B. Dodd and R. Campbell (Lawrence Erlbaum, London, UK), pp. 97–113.
Rosenblum, L. D., Miller, R. M., and Sanchez, K. (**2007**). "Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects," Psychol. Sci. **18**, 392–396.
Rosenblum, L. D., Yakel, D. A., and Green, K. P. (**2000**). "Face and mouth inversion effects on visual and audiovisual speech perception," J. Exp. Psychol. Hum. Percept. Perform. **26**, 806–819.
Scharenborg, O., Mitterer, H., and McQueen, J. M. (**2011**). "Perceptual learning of liquids," in *Proceedings of Interspeech: 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, pp. 149–151.
Sjerps, M. J., and McQueen, J. M. (**2010**). "The bounds on flexibility in speech perception," J. Exp. Psychol. Hum. Percept. Perform. **36**, 195–211.
Spehar, B. P., Tye-Murray, N., and Sommers, M. S. (**2008**). "Intra-versus intermodal integration in young and older adults," J. Acoust. Soc. Am. **123**, 2858–2866.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.
Summerfield, Q. (**1987**). "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading*, edited by B. Dodd and R. Campbell (Lawrence Erlbaum, London, UK), pp. 3–51.
Van Linden, S., and Vroomen, J. (**2007**). "Recalibration of phonetic categories by lipread speech versus lexical information," J. Exp. Psychol. Hum. Percept. Perform. **33**, 1483–1494.
Van Son, N. J. D. M. M., Huiskamp, T. M. I., Bosman, A. J., and Smoorenburg, G. F. (**1994**). "Viseme classifications of Dutch consonants and vowels," J. Acoust. Soc. Am. **96**, 1341–1355.
Walden, B. E., Prosek, R. A., and Worthington, D. W. (**1974**). "Predicting audiovisual consonant recognition performance of hearing-impaired adults," J. Speech Hear. Res. **17**, 270–278.
Yakel, D. A., Rosenblum, L. D., and Fortier, M. A. (**2000**). "Effects of talker variability on speechreading," Percept. Psychophys. **62**, 1405–1412.
Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (**1998**). "Quantitative association of vocal-tract and facial behavior," Speech Commun. **26**, 23–43.