Title

Experience with a talker can transfer across modalities to facilitate lipreading

Permalink
https://escholarship.org/uc/item/3jd724qb

Journal
Attention, Perception, & Psychophysics, 75(7)

ISSN
1943-3921

Authors
Sanchez, Kauyumari
Dias, James W
Rosenblum, Lawrence D

Publication Date
2013-10-01

DOI
10.3758/s13414-013-0534-x

Peer reviewed

Published in final edited form as:

Atten Percept Psychophys. 2013 October ; 75(7): . doi:10.3758/s13414-013-0534-x.

Experience with a Talker Can Transfer Across Modalities to Facilitate Lipreading

Kauyumari Sanchez, **James W. Dias**, and **Lawrence D. Rosenblum**
University of California, Riverside

Abstract

Rosenblum, Miller, and Sanchez (2007) found that participants first trained to lipread a particular talker were then better able to perceive the *auditory* speech of that same talker, compared to that of a novel talker. This suggests that the talker experience a perceiver gains in one sensory modality can be transferred to another modality to make that speech easier to perceive. An experiment was conducted to examine whether this cross-sensory transfer of talker experience could occur: 1) from auditory to lipread speech; 2) with subjects not screened for adequate lipreading skill; 3) when both a familiar and unfamiliar talker are presented during lipreading; and 4) for both old (presentation set) and new words. Subjects were first asked to identify a set of words from a talker. They were then asked to perform a lipreading task from two faces, one of which was of the same talker they heard in the first phase of the experiment. Results revealed that subjects who lipread from the same talker they had heard performed better than those who lipead a different talker, regardless of whether the words were old or new. These results add further evidence that learning of amodal talker information can facilitate speech perception across modalities and also suggest that this information is not restricted to previously heard words.

Familiarity with a talker's speech can facilitate perception of that speech in myriad settings (see Nygaard, 2005, for a review). For example, listeners are better at understanding the speech of those with whom they are familiar in sub-optimal conditions such as background noise (Bradlow & Pisoni, 1999; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994; Yonan & Sommers, 2000). Talker-familiarity facilitation has also been observed for both recognition and implicit memory tasks (e.g., Bradlow, Nygaard, & Pisoni, 1999; Church & Schacter, 1994; Palmeri, Goldinger, & Pisoni, 1993).

It is now known however, that speech is more than an auditory skill. Visual speech perception, or lipreading, is a part of everyone's speech skills, regardless of one's hearing. Research shows that perceivers use visible articulatory movements of the face to enhance noisy or accented speech, and when acquiring a new or first language (see Rosenblum, 2005 for a review). Auditory and visual speech is combined in perception. The McGurk effect, an illusion based on discrepant but synchronous audio and visual syllables, demonstrates that visible speech is incorporated automatically and thoroughly enough to influence what perceivers report *hearing* (McGurk & MacDonald, 1976).

Returning to talker facilitation of speech, visual speech examinations have found similar effects on perception as auditory-only studies. For example, familiar faces can be lipread more easily than unfamiliar faces (Lander & Davies, 2008).

Correspondences should be addressed to Lawrence D. Rosenblum, Department of Psychology, 900 University Ave., University of California, Riverside, CA 92521.

What are the salient talker dimensions with which observers become familiar to facilitate speech perception? In the case of auditory speech, it could be that listeners become familiar with the vocal quality, timbre and general sound of a talker's voice allowing them to better hear the talker's speech when, say, imbedded in noise. For visual speech, it may be that perceivers become familiar with the facial features, and visible mouth characteristics allowing them to lipread more easily.

However, another possibility exists for the information supporting talker facilitation effects. It could be that perceivers become familiar with a talker's general articulatory style that can be reflected in both the auditory and visual signal. This *idiolectic* information would be amodal in being available in both modalities. This possibility motivated us to pursue an examination of talker familiarity effects in a cross-sensory context (Rosenblum et al., 2007). We reasoned that if at least part of the information perceivers use for familiar talker facilitation takes an amodal, articulatory form, then it should be the case that gaining familiarity with a talker in one modality will transfer across modalities to facilitate perception of speech in the other modality.

We tested this prediction by first giving normal hearing subjects, screened for moderate lipreading proficiency, one hour of experience lipreading sentences from one of two talkers. Subjects were then given an auditory speech-in-noise task hearing either the same talker from which they had just lipread, or a different talker. (Subjects were unaware of the talker being the same or different across the tasks.) Three different signal-to-noise ratios were used for the speech-in-noise test. Results revealed that subjects who lipread and heard speech from the same talker performed better on all levels of the speech-in-noise task than did subjects who lipread from one talker, and then heard speech from a different talker.

We interpret these results as evidence that lipreading from a talker provides experience, in part, with amodal talker-specific articulatory (idiolectic) properties which can be transferred cross-modally to facilitate recovery of that talker's auditory speech (Rosenblum et al., 2007). This interpretation is consistent with findings, based on sinewave and point-light speech, showing that there is talker information in the isolated articulatory movements available both auditorily and visually (e.g., Remez et al., 1997; Rosenblum et al., 2004; Rosenblum et al., 2002). Our interpretation is also consistent with results showing that there is some shared talker-specific articulatory information available across modalities allowing perceivers to match voices to faces using isolated articulation (e.g., Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b, 2004c; Rosenblum et al., 2006; and see also Cvejic, et al., 2012).

While our results demonstrate an initial instance of cross-modal transfer of talker information, many questions remain. Most obviously, might this transfer work in the opposite direction: from auditory to visual speech? If there is truly amodal talker-specific information that can act to facilitate speech perception, then it should be extractable from auditory speech and then transfer to, and facilitate, visual speech perception.

A second question is whether the cross-modal transfer would hold for subjects not screened for moderate lipreading proficiency, as had the subjects in our original study. The choice to use subjects of moderate lipreading skill in the original experiment was based on an assumption that to endure the 1-hour sentence lipreading phase of the experiment, they would need to experience some success with the task. This moderate lipreading requirement allowed less than 50% of all subjects tested to be included in the experiment. However, if there is amodal articulatory style information available in visible speech, then there is no reason why at least some of this information should not be available to all perceivers, regardless of lipreading skill.

A third question concerns whether cross-modal transfer will occur if, during testing, the ‘familiar’ talker is perceived intermixed with another talker. Our previous study involved a between subjects design for which one subject group heard the same speaker they had just lipread, and a second group heard a different speaker. In both cases, the test phase involved perceiving speech from a single speaker. It is unclear whether having subjects perceive speech from both a familiar and unfamiliar speaker during the test phase would also show the cross-modal transfer effect. It could be that intermixing the speech of two different talkers during test would induce a distraction so as to eliminate any crossmodal talker facilitation.

Another remaining question is whether the extracted facilitating talker information is retained at the word level or/and at another linguistic level. There is reason to think that it may be held at the word level. According to a prominent theory of word recognition, Goldinger's theory of the episodic lexicon (e.g., Goldinger, 1998), when words are heard, lexical information about words and talker-specific information (how it was said and by whom) are stored together in a memory episode. When these episodes are brought forth from long-term memory into working memory, and the information matches the input, the episodes can enhance perception. This theory could explain the talker facilitation of speech observed for both auditory and visual speech perception (although it was originally devised to explain only auditory speech).

Importantly, the episodic lexicon theory states that talker information is stored at the word level. Presumably then, any talker facilitation of speech, whether auditory, visual, or cross-modal, should exist chiefly with the specific words previously perceived from the familiar talker. Other theories of talker influences proffer that talker-specific information can be effective at a sub-lexical level (e.g., Cutler, et al. 2010; Johnson, 1997; Pierrehumbert, 2002), retained with phonemes or features, and generalizable across the lexicon. Clearly, these theories would suggest that as long as some phonetic overlap exists between study and test material, talker facilitation could occur regardless of whether the same words were used in both tests. In fact, these theories are consistent with findings testing accented (e.g., Sidaras, et al. 2009; Wingstedt & Schulman, 1987**)** and distorted speech (vocoded; compressed; sinewave; e.g., Hervais-Adelman, et al. 2008; and see Banai & Lavner, 2012 and Bent, et al. 2011 for reviews) showing that perceptual learning and adaptation can generalize to new lexical items (see also Bradlow & Pisoni, 1999; McQueen, Cutler, & Norris, 2006; Nielsen, 2011; Nygaard & Pisoni, 1998; Nygaard, et al. 1994; Yonan & Sommers, 2000). While these theories chiefly address the influence of talker information for *auditory* speech, they can easily be extrapolated to predict sublexical effects for visual speech and crossmodal talker effects.

In our previous cross-sensory transfer experiment, we could not easily examine the question of whether the facilitative talker information was maintained at the lexical or/and sublexical level . In that experiment, separate sets of sentences were used for the lipreading presentation phase and the speech-in-noise test phase of the experiment. While there was likely some lexical overlap between the two sets, the overlapping words were not counterbalanced in important characteristics such as lexical frequency (known to influence talker-facilitation effects such that less frequent words show greater talker influences, e.g., Goldinger, 1998). Thus, it is yet unknown whether cross-modal transfer of talker facilitation would occur chiefly at the word level, as the episodic lexicon theory would seem to predict, or at a sublexical level, as other theories would predict.

The current investigation attempts to address the four preceding questions by testing (1) auditory-to-visual transfer with (2) subjects not screened for lipreading competence, who are (3) asked to lipread both a familiar and unfamiliar talker articulating (4) both words heard and not heard in the presentation set. To accomplish this, a number of major modifications were made from our original test. Subjects were asked to first identify heard words, all spoken by one of two talkers. These words were varied across four levels of lexical frequency, and were each presented five times during the presentation set – both manipulations used in assessing the old/new word question (see below). After this presentation set, subjects were asked to lipread words from two talkers, one from which they had (unknowingly) just heard. The lipread word set was composed of 'old' words subjects had just heard, as well as new words. Finally, because of the difficulty of lipreading words for non-screened subjects, lipreading performance was evaluated by determining the percentage of phonemes correctly lipread.

If speech-facilitating talker information can be transferred from auditory to visual speech perception, and can do so with non-prescreened subjects asked to lipread both a 'familiar' and unfamiliar talker, then subjects should be better at lipreading from the (familiar) talker they had experience hearing. Further, if this cross-modal talker information is maintained at the word level, then this familiar talker advantage should be greater for old than new words.

Method

Participants

Thirty-two participants (7 males, 25 females) were recruited from undergraduate psychology classes at the University of California, Riverside. Participants completed the study in partial fulfillment of course credit. All participants were native English speaking with normal hearing and normal, or corrected-to-normal, vision.

Materials and Stimuli

Two female models were audio-video recorded uttering 160 words (see Goldinger & Azuma, 2004), varying in lexical frequency (Kucera & Francis, 1967). Both models were native to southern California. The models were recorded reading words from a teleprompter at a rate of one word per second. After each word utterance was saved on a computer, the audio and visual streams were separated and each saved as their own unique token. The audio tokens were then amplitude normalized and the video tokens were edited to show only the lower half of the models' face (bottom of the throat to upper cheeks, just below the eyes) silently articulating a word (Rosenblum et al., 2007).

Design and Procedure

This study used a 2 (talker-type) by 2 (word-type) by 4 (lexical frequency) within subjects design. Talker-type (same [trained], different [novel]), word-type (old, new) and lexical frequency (high, medium-high, medium-low, low) were all measured as within subjects variables.

Subjects completed the experiment in a sound-attenuated booth. Subjects first participated in a training phase, for which they were asked to identify each of the utterances they heard spoken from a single talker. Subjects were asked to say each word aloud, quickly and clearly, following the auditory presentation of that word. One-hundred and six words, with all four word frequencies being equally represented, were presented in the training phase. Several different combinations of the word lists were created from the full list of 160 words, to counterbalance the old and new words said by each talker in the test phase for different participants. Also, half of the subjects heard words spoken by one of the two talkers, the other half heard words spoken by the other talker.

In this training phase, each word was pseudo-randomly presented five times each for a total of 530 utterances. Motivation for word repetition in this task was two-fold. First, word repetition was used to create several memory episodes containing word and talker information. The episodic encoding theory predicts and has found that increasing repetitions facilitates recognition memory (e.g. Palmeri et al., 1993; Goldinger, 1998). This could allow for a stronger test of the old/new word prediction. The second reason for word repetition was to provide the participant with sufficient exposure to the model's speech. The Rosenblum et al. (2007) study trained participants for roughly an hour with sentence-length tokens, providing participants with ample exposure to the models' speech. Word repetition in the current experiment provided roughly 45 minutes of model exposure. All subjects were able to identify all words in the training phase.

The test phase consisted of a word identification task for visual-only stimuli. Participants were asked to lip-read word tokens from two talkers, one of whom was (unbeknownst to subjects) the same talker whose voice was heard in the training phase. The test phase consisted of 106 old words and 54 new words presented randomly. The old and new words were equally distributed between the talkers, providing 53 old words and 27 new words each. Each trial consisted of two presentations of the same visual-only word token. Participants were instructed to identify (as best they could) each word by uttering their response aloud after the second presentation. Research assistants outside the sound booth recorded responses, to be used for later coding.

Results

As mentioned, lipreading performance was evaluated using proportion phonemes correct in lieu of words correct, due to the difficulty of unpracticed subjects in lip-reading words. (On average, subjects lipread only 16% of the 160 test words correctly.) Additionally, phonemes rather than visemes correct were calculated, because the literature is not at a consensus regarding which segments are truly visible (e.g. Lidestam & Beskow, 2006). Target words and participant responses (which were written down as text by an experimenter – see above) were transcribed from orthographic to phonemic strings using an automatic transcription procedure based on the built-in text-to-speech feature in Mac OS X (Dias & Lachs, in review). These phonetic transcriptions were also checked by hand. Based on these transcriptions, mean percent accuracy was calculated for the phonemes of the target lip-read words and was used in a 2 by 2 by 4 repeated measures Analysis of Variance (ANOVA).

With respect to familiar vs. unfamiliar talker, a significant main effect was found, $F(1, 31) = 4.508$, $p < .05$, $\eta_p^2 = .127$. Participants were better at lipreading the talker with whom they had auditory familiarity ($M = .455$, $SE = .017$), compared to the unfamiliar talker ($M = .429$, $SE = .014$) (see Figure 1). There was also a significant main effect for word-type, $F(1, 31) = 29.653$, $p < .001$, $\eta_p^2 = .489$, such that participants were better at lip-reading (old) words they had heard in the training phase ($M = .468$, $SE = .017$) than new words ($M = .415$, $SE = .013$). This is consistent with previous findings showing that prior exposure to target words facilitates lip-reading of those words (e.g., Miller et al., 2010; Sanchez, 2011).

Lexical frequency also resulted in a significant main effect, $F(3, 29) = 27.291$, $p < .001$, $\eta_p^2 = .468$. Participants were increasingly better at lip-reading as lexical frequency increased (Low M =.370, $SE = .012$; Medium-Low $M = .455$, $SE = .015$; Medium-High $M = .464$, $SE = .018$; High M =.478, $SE = .019$). Bonferroni corrected probes were conducted to determine which of the lexical frequency levels were significantly different from each other. These tests revealed that participants were no more accurate identifying segments in high frequency words ($M = .616$, $SE = .021$) than in medium-high frequency words ($M = .619$, $SE = .023$), $t(61) = -0.322$, $p = .749$, $r = 0.041$. However, participants did identify more segments in high frequency words than in medium-low frequency words (*M* = .583, *SE* = .020), *t*(61) = 3.490, *p* < .001, *r* = .408, and than in low frequency words (*M* = .499, *SE* = .019), *t*(61) = 11.880, *p* < .001, *r* = .836. Participants also identified more segments in medium-high frequency words than in medium-low frequency words, *t*(61) = 3.959, *p* < .001, *r* = .452, and in low frequency words, *t*(61) = 12.199, *p* < .001, *r* .842. Finally, participants identified more segments in medium-low frequency words than in low frequency words, *t*(61) = 8.069, *p* < .001, *r* = .719. Overall, these results are consistent with evidence showing that high frequency words are generally easier to lip-read than low frequency words (Auer, Bernstein, & Tucker, 2000).

T-tests were conducted to determine whether either of the talkers in the stimulus set was easier to lipread, overall (within subjects test), or whether there was a difference in overall lipreading performance depending on which of the talkers was heard in the presentation phase (between subjects test). Neither of these tests showed a significant difference (p>.05) between performance with these talker factors, suggesting that there was no inherent difference between the two stimulus talkers relevant to the critical findings.

Finally, no significant interactions were found between any of the manipulated variables.

Discussion

In revealing a familiar talker effect, these results extend Rosenblum et al.'s (2007) findings by demonstrating that talker information can transfer cross-modally from auditory speech experience to facilitate lipreading performance. The results also show that this transfer of talker facilitation can occur for individuals not pre-screened for lipreading skill suggesting that the talker information that can facilitate speech perception crossmodally is not only useful to subjects with moderate to good lipreading skill. Second, the results show that talker facilitation can transfer across modalities even when perceivers must contend with the potential distraction of lipreading from two intermixed talkers. While the magnitude of the familiar talker effect may seem relatively small, it is comparable to the magnitude of the effect observed by Rosenblum et al. (2007), and to auditory talker facilitation effects (e.g., Nygaard & Pisoni, 1998; Nygaard et al., 1994; Yonan & Sommers, 2000).

Participants were better at lipreading from their auditorily familiar talker regardless of whether the words were old or new. This finding would seem counter to an episodic lexicon theory, which proffers that the facilitating talker information would be contained in memory stores of *lexical* episodes. It seems unlikely that the lack of interaction is simply a by-product of the phoneme level response analysis used for the experiment. This analysis was, after all, able to reveal significant word-relevant effects including lexical frequency and an *overall* old/new word influence.

Instead, our findings would seem supportive of theories proffering that talker information can also reside at sublexical levels allowing for facilitation of speech perception which is generalized across the lexicon (e.g., Cutler, et al. 2010; Johnson, 1997; Pierrehumbert, 2002). As stated, these theories are consistent with findings on accented and distorted speech, showing that learning to better perceive these forms generalizes to new words (see also Bradlow & Pisoni, 1999; McQueen, Cutler, & Norris, 2006; Nielsen, 2011; Nygaard & Pisoni, 1998; Nygaard, et al. 1994; Yonan & Sommers, 2000). In finding talker effects generalized to new words, the current findings would also seem consistent with these theories, despite involving crossmodal effects. As stated, it is likely that these theories could be modified relatively easily to address the crossmodal effects reported here and in our earlier study (Rosenblum, et al. 2007).

Informational Support for Crossmodal Talker Facilitation Effects

The fact that facilitative talker information can transfer bi-directionally across modalities adds evidence that the relevant auditory and visual information can jointly specify some underlying amodal, articulatory properties that are talker-specific. The question arises of what, exactly, these identifying articulatory properties are, and how might they be specified auditorily and visually?

While neither the current experiment, nor that of Rosenblum, et al. (2007) was designed to examine the specific information for crossmodal transfer, speculation can be provided based on related research. For example, the audio and visual information supportive of crossmodal *talker matching* has been examined (e.g., Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b, 2004c; Rosenblum et al., 2006; and see also Cvejic, et al., 2012). As stated, some of this research has shown that the isolated time-varying information contained in sinewave auditory speech and point-light visual speech is sufficient to support matches (Lachs & Pisoni, 2004b; Rosenblum, et al. 2006; but see Mavica & Barenholtz, 2013). Other research has shown that the same auditory dimensions allowing for cross-modal talker matches are strongly salient for phonetic perception (Lachs & Pisoni, 2004a). Lachs and Pisoni (2004a) tested 14 different acoustic manipulations (e.g., spectral inversion and non-linear scaling) and found a very high correlation between those that reduced crossmodal talker matching and those that inhibited word recognition. In general, these authors found that as long as relative spectral and temporal patterning is maintained (e.g., ratio of formant spacing is maintained over time), other dimensions —such as absolute temporal patterning— can be distorted without suppression of matching and word identification performance. From these findings, Lachs and Pisoni conclude that “indexical and linguistic information are carried in parallel components of the acoustic signal and do not appear to be dissociable from one another” (p.392-393).

This interpretation is consistent with the explanation provided by Remez and colleagues (1997) in finding that the isolated phonetic information retained in sinewave speech could jointly support phonetic and talker recognition. Remez, et al. (1997) speculate that this information may be contained in segment-level coarticulatory assimilation considered to be largely idiosyncratic to a talker's speaking style (idiolect) (see also Amerman & Daniloff, 1977; Bladon & Al-Bamerni, 1976; Sheffert et al., 2002). If there is overlap in phonetic and talker-specific information, then learning to recognize a talker could involve, partly, learning to attend to information that also supports better perception of that talker's speech.

In fact, some of the earliest research on talker familiarity effects has shown that the act of learning to *recognize* a talker facilitates perception that talker's speech (see Nygaard, 2005, for a review). In fact, the seminal study by Nygaard and her colleagues (Nygaard, et al. 1994) included a presentation phase for which listeners were tasked with talker learning and then a ‘surprise’ test phase where listeners performed a phonetic recognition task (of old and new talkers). Thus, the extant research suggests that there is overlapping information underlying talker and speech learning, at least in the auditory domain. The research also suggests that that the relevant auditory information can take a form of relational temporal-spectral structure (e.g., Lachs & Pisoni, 2004b) that could specify the coarticulatory assimilation specific to a talker's idiolect (Remez, et al. 1996).

Turning to visual speech, it could be that an analogous form of information exists in the optic signal. If coarticulatory assimilation turns out to be a talker property that jointly supports talker and speech identification, then there is no reason that this distal event could not be specified both acoustically *and optically* —at least to some degree. In fact, there is evidence that visual speech information can be similar in form to acoustic speech information for purposes of both phonetic and talker perception (see Rosenblum, 2005; 2008 for a review). Point-light visible speech is analogous to sinewave speech in showing that time-varying information for articulation can be used for both speech and speaker identification (Rosenblum, et al. 1996; Rosenblum, et al. 2002; 2006), as well as crossmodal talker matching (Lachs & Pisoni, 2004; Rosenblum, et al. 2006). In fact, both point-light and sinewave speech have been considered kinematic information that can specify a common underlying articulatory dynamics (Lachs & Pisoni, 2004b; Rosenblum, 2005). With regard to the possibility of overlapping visual information for speech and speaker recognition, Lachs and Pisoni (2004c) found that similar to their observations of auditory speech, manipulations which inhibited talker crossmodal matching (e.g., temporal reversal) also affected visual phonetic perception.

Returning to the current findings, it could be that during training, our subjects' attempts to recognize the auditory speech of a talker allowed them to become familiar with that talker's fine-grain, phonetically-relevant idiolectic dimensions (e.g., coarticulatory assimilation) as specified in the acoustic signal. This familiarity may then have allowed our subjects to better recover these same idiolectic dimensions of that talker as it was specified visually during the test phase of the experiment. A similar explanation can be applied to our previous crossmodal talker familiarity findings (Rosenblum, et al. 2007), except with the auditory and visual speech phases reversed.

Interestingly, there was nothing conspicuous distinguishing our two talkers' auditory or visual speaking styles, based on the casual impressions of the authors and a number of subject comments. In fact, the talkers were chosen to be superficially similar so that subjects would be unaware of crossmodal correspondence of identity (the same was true of the talkers in the Rosenblum, et al, 2007 study). Thus, whatever idiolectic information our subjects learned that allowed for crossmodal transfer, it was likely at a very nuanced, fine-grain level – a conclusion consistent with the conjecture that dimensions such as coarticulatory assimilation may be useful.

Future research can determine whether fine-grain dimensions articulatory dimensions such as coarticulatory assimilation are relevant for talker familiarity effects and whether these amodal dimensions can be specified in both the acoustic and optic signal. Based on the extant research (e.g., Lachs & Pisoni, 2004b), it can be speculated that these dimensions will be specified in the relational spectral-temporal patterning of the acoustic signal and, possibly, some analogue patterning in the optic signal.

The chief finding of cross-modal transfer of implicit perceptual experience is also consistent with some recent non-speech examples. These examples have shown that implicit perceptual experience obtained in one modality can transfer to another modality in the form of perceptual aftereffects. These demonstrations include auditory and visual perception of rate (Levitan et al., 2012), perceived direction of visual and tactile stimulation (Konkle et al., 2009), and auditorily and visually perceived approach direction (Kitagawa & Ichihara, 2002). Each of these cross-modal effects have been explained by appeal to how perceptual experience can influence some abstract (i.e., amodal) perceptual properties that can transfer across modalities. Future research on multisensory speech and non-speech can examine the degree to which such explanations are correct.

Acknowledgments

This research was supported by NIDCD Grant 1R01DC008957-01.

References

Amerman JD, Daniloff RG. Aspects of lingual coarticulation. Journal of Phonetics. 1977; 5:107–113.

Auer ET, Bernstein LE, Tucker PE. Is subjective word familiarity a meter of ambient language? A natural experiment on the effects of perceptual experience. Memory & Cognition. 2000; 28:789–797. [PubMed: 10983453]

Banai K, Lavner Y. Perceptual Learning of Time-Compressed Speech: More than Rapid Adaptation. PLOS One. 2012; 7:1–9.

Bent T, Loebach JL, Phillips L, Pisoni DB. Perceptual Adaptation to Sinewave-Vocoded Speech Across Languages. Journal of Experimental Psychology-Human Perception and Performance. 2011; 37:1607–1616. [PubMed: 21688936]

Bladon RAW, A1-Bamerni A. Coarticulation resistance in English IlL. Journal of Phonetics. 1976; 4:137–150.

Bradlow AR, Pisoni DB. Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. Journal of the Acoustical Society of America. 1999; 106:2074–2085. [PubMed: 10530030]

Church BA, Schacter DL. Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1994; 20(3):521–533.

Cutler A, Eisner F, McQueen JM, Norris D. How abstract phonemic categories are necessary for coping with speaker-related variation. Labphon 10. 2010; 10:91–111.

Cvejic E, Kim J, Davis C. Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody. Cognition. 2012; 122:442–453. [PubMed: 22196745]

Dias JW, Lachs L. The role of cross-modal speech information in audiovisual enhancement. Perception. in review.

Goldinger SD. Words and voices: Episodic traces in spoken word identification and recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1996; 22:1166–1183.

Goldinger SD. Echoes of echoes? An episodic theory of lexical access. Psychological Review. 1998; 105:251–279. [PubMed: 9577239]

Goldinger SD, Azuma T. Episodic memory reflected in printed word naming. Psychonomic Bulletin & Review. 2004; 11:716–722. [PubMed: 15581123]

Goldinger SD, Kleider HM, Shelley E. The marriage of perception and memory: Creating two-way illusions with words and voices. Memory & Cognition. 1999; 27:328–338. [PubMed: 10226442]

Hervais-Adelman A, Davis MH, Johnsrude IS, Carlyon RP. Perceptual learning of noise vocoded words: Effects of feedback and lexicality. Journal of Experimental Psychology: Human Perception and Performance. 2008; 34:460–474. [PubMed: 18377182]

Johnson, K. Speech perception without speaker normalization: An exemplar model.. In: Johnson, K.; Mullennix, JW., editors. Talker Variability in Speech Processing. Academic Press; San Diego: 1997. p. 145-66.

Kamachi M, Hill H, Lander K, Vatikiotia-Bateson E. Putting the face to the voice: Matching identity across modality. Current Biology. 2003; 13:1709–1714. [PubMed: 14521837]

Kitagawa N, Ichihara S. Hearing visual motion in depth. Nature. 2002; 416:172–174. [PubMed: 11894093]

Konkle T, Wang Q, Hayward V, Moore CI. Motion Aftereffects Transfer between Touch and Vision. Current Biology. 2009; 19:745–750. [PubMed: 19361996]

Kucera, H.; Francis, W. Computational analysis of present-day American English. Brown University Press; Providence, RI: 1967.

Lachs L, Pisoni DB. Specification of cross-modal source information in isolated kinematic displays of speech. Journal of the Acoustical Society of America. 2004a; 107:507–516. [PubMed: 15296010]

Lachs L, Pisoni DB. Crossmodal source information and spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance. 2004b; 30:378–396. [PubMed: 15053696]

Lachs L, Pisoni DB. Crossmodal source identification in speech perception. Ecological Psychology. 2004c; 16:159–187. [PubMed: 21544262]

Lander K, Davies R. Does familiarity influence speechreadability? The Quarterly Journal Of Experimental Psychology. 2008; 61:961–967. [PubMed: 18570135]

Levitan, CA.; Ban, YA.; Stiles, NB.; Shimojo, S. Cross-modal transfer without concurrent stimulation: a challenge to a hidden assumption.. Poster presented at the 12th Annual Meeting of the Vision Sciences Society; May 15; 2012.

Lidestam B, Beskow J. Visual phonemic ambiguity and speechreading. Journal of Speech, Language, & Hearing Research. 2006; 49:835–847.

Mavica LQ, Barenholtz E. Matching Voice and Face Identity From Static Images. Journal of Experimental Psychology: Human Perception & Performance. 2013; 39:307–312. [PubMed: 23276114]

McQueen JM, Cutler A, Norris D. Phonological abstraction in the mental lexicon. Cognitive Science. 2006; 30:1113–1126. [PubMed: 21702849]

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264:746–748. [PubMed: 1012311]

Miller RM, Sanchez K, Rosenblum LD. Alignment to visual speech information. Attention, Perception, & Psychophysics. 2010; 72:1614–1625.

Nielsen KY. Specificity and abstractness of VOT imitation. Journal of Phonetics. 2011; 39:132–142.

Nygaard, LC. The integration of linguistic and non-linguistic properties of speech.. In: Pisoni, D.; Remez, R., editors. Hand-book of speech perception. Blackwell; Malden, MA: 2005. p. 390-414.

Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. Perception & Psychophysics. 1998; 60:355–376. [PubMed: 9599989]

Nygaard LC, Sommers MS, Pisoni DB. Speech perception as a talker-contingent process. Psychological Science. 1994; 5:42–46. [PubMed: 21526138]

Palmeri TJ, Goldinger SD, Pisoni DB. Episodic encoding of voiceattributes and recognition memory for spoken words. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1993; 19:309–328.

Pierrehumbert, JB. Word-specific phonetics.. In: Gussenhoven, C.; Warner, N., editors. Laboratory phonology. Vol. VII. Mouton de Gruyter; Berlin: 2002. p. 101-139.

Remez RE, Fellowes JM, Rubin PE. Talker identification based on phonetic information. Journal of Experimental Psychology: Human Perception & Performance. 1997; 23:651–666. [PubMed: 9180039]

Rosenblum, LD. The primacy of multimodal speech perception.. In: Pisoni, D.; Remez, R., editors. Handbook of Speech Perception. Blackwell; Malden, MA: 2005. p. 51-78.

Rosenblum L, Miller R, Sanchez K. Lipread me now, hear me better later: Cross-modal transfer of talker familiarity effects. Psychological Science. 2007; 18:392–396. [PubMed: 17576277]

Rosenblum LD, Smith NM, Hale S, Lee J. Hearing a face: Cross-modal speaker matching using isolated visible speech. Perception and Psychophysics. 2006; 6:84–93. [PubMed: 16617832]

Rosenblum LD, Yakel DA, Baseer N, Panchal A, Nordarse BC, Niehus RP. Visual speech information for face recognition. Perception & Psychophysics. 2002; 64(2):220–229. [PubMed: 12013377]

Sheffert SM, Pisoni DB, Fellowes JM, Remez RE. Learning to recognize talkers from natural, sinewave, and reverse speech samples. Journal of Experimental Psychology: Human Perception & Performance. 2002; 28:1447–1469. [PubMed: 12542137]

Sidaras SK, Alexander JED, Nygaard LC. Perceptual learning of systematic variation in Spanish-accented speech. Journal of the Acoustical Society of America. 2009; 125:3306–3316. [PubMed: 19425672]

Wingstedt, M.; Schulman, R. Comprehension of foreign accents.. In: Dressler, W.; Luschutzky, H.; Pfeiffer, O.; Rennison, J., editors. Phonologica 1984. Cambridge U.P.; Cambridge: 1987. p. 339-345.

Yonan CA, Sommers MS. The effects of talker familiarity on spoken word identification in younger and older listeners. Psychology and Aging. 2000; 15:88–99. [PubMed: 10755292]

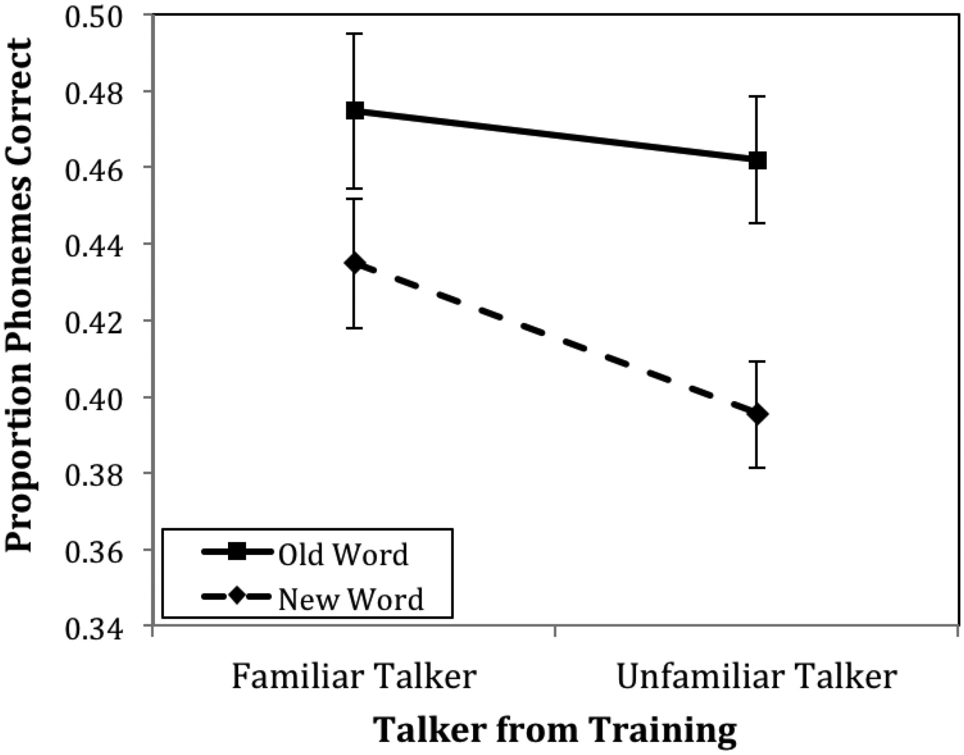

Figure 1.
Mean proportions of correctly identified phonemes from old and new words for familiar and unfamiliar talkers. Error bars represented standard errors.