Attention, Perception, & Psychophysics (2021) 83:415–434
https://doi.org/10.3758/s13414-020-02141-9

Cross-modal transfer of talker-identity learning

Dominique Simmons[1] · Josh Dorsi[1] · James W. Dias[1] · Lawrence D. Rosenblum[1]

Accepted: 4 September 2020 / Published online: 20 October 2020

Abstract
A speech signal carries information about meaning and about the talker conveying that meaning. It is now known that these two dimensions are related. There is evidence that gaining experience with a particular talker in one modality not only facilitates better phonetic perception in that modality, but also transfers across modalities to allow better phonetic perception in the other. This finding suggests that experience with a talker provides familiarity with some amodal properties of their articulation such that the experience can be shared across modalities. The present study investigates if experience with talker-specific articulatory information can also support cross-modal *talker* learning. In Experiment 1 we show that participants can learn to identify ten novel talkers from point-light and sinewave speech, expanding on prior work. Point-light and sinewave speech also supported similar talker identification accuracies, and similar patterns of talker confusions were found across stimulus types. Experiment 2 showed these stimuli could also support cross-modal talker matching, further expanding on prior work. Finally, in Experiment 3 we show that learning to identify talkers in one modality (visual-only point-light speech) facilitates learning of those same talkers in another modality (auditory-only sinewave speech). These results suggest that some of the information for talker identity takes a modality-independent form.

Keywords Multisensory processing · Speech perception · Face perception

Introduction

The last 20 years have shown tremendous growth in the research concerning the cross-modal transfer of sensory experience. For example, it has been shown that motion aftereffects can be transferred across the visual and tactile modalities (Konkle, Wang, Hayward, & Moore, 2009). Relatedly, there is evidence that stimulus-timing information can be transferred between the auditory and visual modalities (Levitan, Ban, Stiles, & Shimojo, 2015). These low-level perceptual aftereffects are consistent with what has been reported for more complex stimuli. There is evidence, for example, of haptic-visual cross-modal transfer of facial expression (Matsumiya, 2013). There is also evidence that substantial cross-modal learning can occur implicitly and with unattended aspects of stimulation (e.g., Seitz & Watanabe, 2005).

Within the realm of speech there is evidence that bimodal audiovisual experience results in improved auditory-only *talker identification* (the *bimodal training effect*; e.g., von Kriegstein & Giraud, 2006). While these effects refer specifically to audiovisual talker learning effects, an important finding associated with them is the functional coupling between brain areas associated with face and voice processing (Blank, Anwander, & von Kriegstein, 2011; Schall & von Kriegstein, 2014; von Kriegstein & Giraud, 2006; von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005). Importantly, this cross-modal transfer of talker identity information is not merely the result of associative experience: Audiovisual experience with less physically grounded stimuli, such as written names, does not improve auditory talker identification (von Kriegstein & Giraud, 2006). Moreover, there is evidence that hearing a word will facilitate later visual-only identification of that word (van der Zande et al., 2014a). Similarly, experience with a talker in one modality can facilitate the perception of that talker's *speech* in a different modality (e.g., Rosenblum, Miller, & Sanchez, 2007; Sanchez et al., 2013). Based on this literature it seems that information about both speech and talker is cross-modally available. These findings prompt the central question of this investigation: can experience learning to unimodally *identify a talker* in one modality be shared across modalities?

In fact, there is strong evidence that talker identification information is cross-modally available (e.g., Kamachi, Hill,

✉ Lawrence D. Rosenblum
lawrence.rosenblum@ucr.edu

[1] Department of Psychology, University of California, Riverside, Riverside, CA 92521, USA

Lander, & Vatikiotis-Bateson, 2003; Lachs, & Pisoni, 2004a; Lachs & Pisoni, 2004b; Lachs & Pisoni, 2004c; Rosenblum, Smith, Nichols, Hale, & Lee, 2006). These findings show that observers can match speaking faces to voices even when different words are spoken in each modality. This work shows that cross-modal talker matching is possible even when face images are reduced to articulatory information alone. Similar results have been found using highly reduced auditory stimuli designed to isolated "phonetic" information (Lachs & Pisoni, 2004c). Authors of these reports suggest that cross-modal talker matching could be based on the extraction of talker-specific articulatory style information available across modalities.

Substantial research shows that talker information is carried in the phonetic realizations of speech (Remez, Fellowes, & Rubin, 1997). Sinewave speech re-synthesis has been a valuable tool in this research. Sinewave speech (Remez et al., 1987) allows researchers to exclude all of the classic information thought to indicate talkers (e.g., fundamental frequency; the spectral structure underlying breathiness). These sinewave speech signals are composed of three simple sinewaves whose trajectories mimic the center formant frequencies of the first two or three formants of a recorded speech signal. Despite sounding like whistles, listeners can extract phonetic information from these signals (Remez et al., 1987; Remez, Rubin, Pisoni, & Carrell, 1981).

Relevant to the present investigation, Remez and his colleagues have found that these signals can also convey talker information, allowing recognition of the talkers from which the signals are derived. Remez and his colleagues (1997) argue that talker recognition is possible with sinewave speech, because while it lacks typical talker-specific identifiers, it does contain talker-specific phonetic information conveyed through the spectral-temporal dynamics of the acoustic signal (see also Fellowes, Remez, & Rubin, 1997). Importantly, listeners can also learn to identify novel talkers through training with sinewave speech (Sheffert, Pisoni, Fellowes, & Remez, 2002).

Other research suggests that the same may be true of visual speech perception. This research has made use of a facial point-light technique in which small fluorescent dots are placed on the lips, teeth, and face of a talker (Rosenblum, Johnson, & Saldana, 1996; Rosenblum & Saldana, 1996). When filmed in the dark, the resultant video image shows only moving dots against a black background, thereby removing the typical facial features thought necessary to identify faces. Without movement, these images cannot even be identified as faces, let alone individual talkers. However, the articulatory information conveyed through the movements of the dots in these point-light videos is sufficient to support speech and talker recognition (Rosenblum, Yakel, Baseer, Panchal, Nodarse, & Niehaus, 2002; Rosenblum, Niehus, & Smith, 2007).

More recent research with point-light speech now demonstrates that these stimuli can also support talker *learning* (Jesse & Bartoli, 2018). In a recent study, participants were trained to identify two novel talkers in point-light speech. Participants were provided feedback during a two alternative, talker identification task using point-light sentences. In a post-test, participants were asked to identify point-light sentences of the same talkers, now without feedback. Performance on this post-test revealed broad talker learning. Participants were not only able to correctly identify talkers for trials that presented different recorded utterances of sentences used during training, but could also identify the talkers from fully novel sentences.

Subsequent experiments from this study also revealed that this sort of talker learning could be extended to the more difficult task of identifying four talkers. Furthermore, these researchers found that their talker learning effects were not driven by differences in talker sex, (Jesse & Bartoli, 2018; Jesse & Saba 2017). These results are complementary to the sinewave speech results of Remez and his colleagues (i.e., Sheffert et al., 2002) who (as noted above) showed that unfamiliar talkers could be learned and subsequently recognized in sinewave speech. These two studies are similar in revealing how isolated phonetic information about talkers can be learned.

What has not been determined is whether learning about specific talkers in one modality can facilitate learning to identify those same talkers in another. As stated above, it has been demonstrated that experience with a talker's auditory speech can improve the understanding of that talker's visual speech (Sanchez et al., 2013). In that study, participants were first asked to identify words from audio-only speech from a single talker. In a subsequent task, participants were asked to identify words from visual-only ("lip-read") speech from either the talker that had been presented in the audio-only block or from a novel talker. Results found that lip-reading was more accurate for talkers whom the participant had previously listened to, despite their never having experience with an audio-visual presentation of those talkers. Similarly, experience with a talker's visual speech facilitated the perception of that talker's auditory-only speech (Rosenblum et al., 2007a). These results indicate that talker-specific information for facilitating phonetic perception can be shared across modalities.

In short, the literature indicates that talker familiarity can cross-modally facilitate the identification of phonetic material (i.e., Rosenblum et al., 2007; Sanchez et al., 2013; van der Zande et al., 2014). There is also evidence that isolated phonetic information in auditory-only (i.e., sinewave speech) and visual-only (i.e., point-light speech) is sufficient for talker identification (i.e., Jesse & Bartoli, 2018; Sheffert et al., 2002). Building on this literature, the research presented here investigated whether this isolated phonetic information supports learning to *identify talkers* in a form that can be shared

cross-modally. For these purposes, we used sinewave (e.g., Remez et al., 1981) and point-light (e.g., Rosenblum et al., 2006) speech to isolate the phonetic information in our audible and visible stimuli in a cross-modal talker-identity learning paradigm.

In Experiment 1, we verified that the information carried by sinewave speech and point-light speech in our particular stimuli allows observers to learn to identify unfamiliar talkers. Experiment 2 showed that participants can match talkers across point-light and sinewave speech with our stimuli. Experiments 1 and 2 used new stimuli, training, and tasks to extend the past research showing talker-identification learning based on articulatory information. Importantly, in using a single set of talkers, these experiments also allowed for the examination of possible common bases of identification across audio and visual contexts, thereby providing an expansion on what has been investigated in similar work (e.g., Jesse & Bartoli, 2018; Sheffert et al., 2002). Finally, Experiment 3 examined whether learning to identify talkers from articulatory information in the visual modality could facilitate learning to identify those same talkers in the auditory modality. If the cross-modally available information for talker learning includes amodal talker-specific information, then talker learning through visual point-light speech should facilitate talker learning with sinewave speech.

Experiment 1: Unimodal talker identity learning

Experiment 1a investigated if observers can learn to identify the voices of unfamiliar talkers from our sinewave stimuli. This experiment conceptually replicates Sheffert et al. (2002) by testing the talker identification of ten sinewave speech talkers, but there are notable differences. The participants in Sheffert et al. (2002) were trained over the course several days until each subject achieved a criterion of 70% talker identification accuracy (with sinewave speech) before being tested on their ability to identify those same talkers producing novel sentences (again as sinewave speech). In contrast, all participants of this experiment completed training during a *single session* and our stimuli consisted of multiple sinewave-transformed utterances of a *single sentence* produced by each of ten different talkers. These changes were instituted based on our previous research showing that less training and less language material can be used in the context of talker matching and talker-facilitated speech perception experiments (e.g., Rosenblum et al., 2002; Rosenblum et al., 2007b; Sanchez et al., 2013). This experiment verified that our sinewave speech stimuli and methods were sufficient to support auditory talker-identity learning before we tested cross-modal talker learning in Experiment 3. Participants were trained to identify ten unfamiliar sinewave speech talkers.

Experiment 1b followed a similar design to Experiment 1a, but used point-light speech in place of sinewave speech. Importantly, the results obtained from Experiments 1a and 1b were then analyzed for similar patterns of confusions to examine a possible common basis of identification across audio and visual stimuli. As stated, this examination expands on what was achieved in previous studies (e.g., Jesse & Bartoli, 2018; Sheffert et al., 2002).

In both experiments, participants were tested using a different set of utterance recordings (of the same sentence) from those they were trained to identify to prevent them from using stimulus idiosyncrasies when identifying talkers. Experiments 1a and 1b served as a valuable assessment of the efficacy of stimuli for Experiments 2 and 3. In addition to this, Experiment 1 also enabled a preliminary test of our central question, whether isolated phonetic information can support cross-modal talker-identity learning through the assessment of talker confusability.

Experiment 1a: Sinewave speech

Method

Participants Comparable to past studies of talker learning (e.g., Jesse & Bartoli, 2018; Sheffert et al., 2002), 19 undergraduates (eight female) from the University of California, Riverside participated in this experiment. Participants received course credit for participation. All participants reported normal or corrected-to-normal hearing and vision. All participants were native speakers of North American English.

Materials Stimuli were recordings of four female and six male native American English talkers, all from the Southern California area. Their ages ranged from 22 to 32 years. Talkers were video-recorded speaking the sentence "The football game is over" nine different times each (corresponding to nine different point-light configurations; see *Materials* section for Experiment 1b below). This sentence was chosen because it has previously been found to be particularly easy to lip-read (e.g., Rosenblum et al., 2002). Talkers were instructed to speak naturally and were not aware of the purpose of the stimuli. Talkers were filmed using a Sony digital video camera (DRC-TRV11) at 30 frames per second. The camera was placed 6 ft away from the talker. (Additional details of the video image are presented under Experiment 1b, below.) A Shure SM57 microphone was placed 1 ft away from the talkers' mouths and was connected to the camera.

Sinewave speech. Audio from these 89 videos (nine videos from nine of the talkers; eight videos from talker F1 who had one video lost; see *Materials* section of Experiment 1b) was extracted using Final Cut Pro software and were normalized.

For each audio token, the center frequencies of the first three formants were extracted using Praat software (Boersma, 2001). These three center frequencies were used to create sinewave replicas of the stimuli (e.g., Remez et al., 1981; Remez et al., 1997). Formant values were adjusted by calculating the mean frequency at 10-ms intervals (Lachs & Pisoni, 2004b). Any apparent differences between the formants of the sinewave speech and the formants of the natural speech were hand corrected. Using these corrected values, three sinusoidal waves were then synthesized using an algorithm in Matlab software (Mathworks, Natick, MA, USA). The finished stimuli preserved time-varying spectro-temporal information while removing natural voice quality (e.g., Sheffert et al., 2002; Remez et al., 1997). Information about talker-specific acoustic dimensions is provided in Table 1 (see also Fig. 1).

Procedure All stimuli were presented using Matlab software (Mathworks, 2010). A ten-key USB numeric keypad was used to record responses. The keypad was labeled with ten names corresponding to each talker. The (fabricated) names used to correspond to each talker were common, monosyllabic, and gender-specific (e.g., Nygaard & Pisoni, 1998; Sheffert et al., 2002).

Participants performed the experiment in a dimmed, sound-attenuated booth. Participants listened to the stimuli using Sony MDR 7506 headphones. Volume was set to a comfortable listening level of 70 db SPL and no participants reported difficulty hearing the sentences.

Familiarization phase. Participants were told that they would learn to identify talkers from auditory stimuli. Participants were also informed that the stimuli they would hear would be different from natural auditory speech and the general features of sinewave speech were explained. For example, sinewave speech was described as sounding like bird chirps or whistles. Participants were told they would be associating each sinewave voice with a particular name and that they would be introduced to the voice-name pairings in the familiarization phase. Participants were told that talkers would always say, "The football game is over" during the entire experiment.

During the familiarization phase, participants were presented with the talker's name followed by an utterance of each talker speaking the sentence "The football game is over" (a blank screen was displayed for the duration of the sinewave utterance). Participants were then presented with the talker's name again.

Participants were presented with two repetitions of one of the utterances from each talker for a total of 20 randomized trials. Participants did not make responses during the familiarization phase.

Training phase. Throughout the training phase, participants were presented with four new sinewave utterances of each talker. Participants were instructed to press a button after each sinewave utterance to indicate which talker they had just heard. Participants indicated their response by pressing a key labeled with the corresponding talker's name on a ten-key numeric keypad. Participants received immediate feedback on each trial. If a response was incorrect, the subject was presented with the correct talker's name. Participants were presented with five repetitions of four new sinewave utterances (except for the one talker who had only eight videos; for this talker, only three different training stimuli were shown, but one of these videos was presented twice as often – see Experiment 1b). This created a total of 200 randomized trials.

Testing phase. The test phase followed the procedure of the training phase but: (1) used the remaining four (of the nine total) sinewave utterances for each talker; and (2) did not provide feedback. On each trial, participants were presented with a sinewave utterance and were then prompted to press a button to identify the talker they had just heard. Participants did not receive feedback following their responses. Participants were presented with five repetitions of the remaining four sinewave utterances of each talker for a total of 200 trials.

Table 1 The duration, intensity, and frequency for the first three formants averaged across all utterances of "The football game is over" for each talker

	Duration (s)	Intensity (db)	F1 (Hz)	F2 (Hz)	F3 (Hz)
F1	2.75 (0.05)	58.53 (0.68)	853.59 (31.61)	1859.45 (43.83)	2808 (105.58)
F2	1.75 (0.05)	60.78 (0.49)	653.67 (7.5)	1689.44 (20.13)	2673.04 (14.31)
F3	2.55 (0.07)	59.5 (0.4)	748.56 (10.5)	1848.8 (28.54)	2871.18 (24.73)
F4	2.13 (0.05)	58.93 (0.59)	688.6 (15.28)	1807.65 (23.59)	2899.06 (8.07)
M1	2.7 (0.08)	58.84 (0.72)	763.68 (16.24)	1744.32 (23.46)	2687.74 (21.39)
M2	1.97 (0.05)	60.92 (0.24)	683.56 (17.8)	1722.23 (21.59)	2641.15 (20.66)
M3	2.04 (0.04)	60.54 (0.48)	646.95 (20.61)	1569.55 (25.72)	2582.07 (30.5)
M4	2.65 (0.09)	58.27 (0.33)	667 (15.21)	1738.72 (31.93)	2737.64 (30.71)
M5	2.43 (0.13)	59.93 (0.86)	692.79 (16.26)	1585.98 (26.49)	2768.38 (47.42)
M6	2.28 (0.07)	58.41 (0.48)	649.29 (7.15)	1818.06 (21.27)	2771.78 (20.5)

Values in the parenthesis are the stadard error

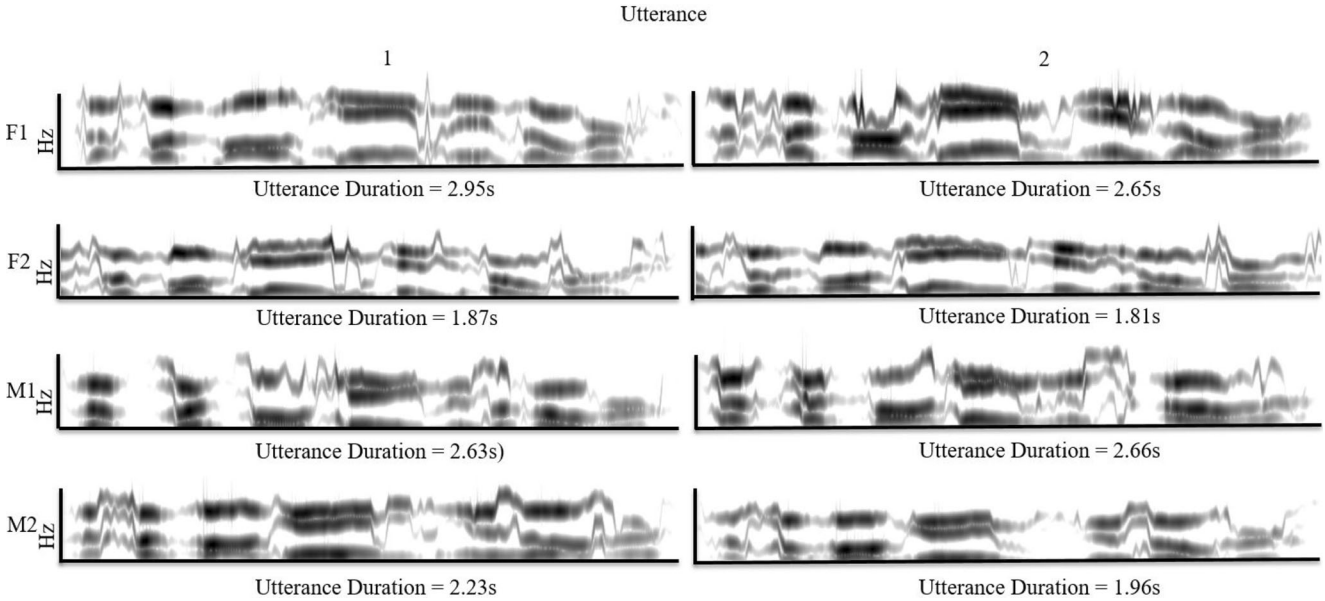

Fig. 1 Spectrograms of two sinewave speech utterances of the sentence "The football game is over" from four (two male) of the talkers used in this study. Note the variability both between different talkers, but also within utterance from a single talker

Results

Training phase

The central question of this experiment is whether listeners could learn to identify talkers from sinewave speech. T-tests were used here and throughout this study to compare talker identification against chance. This statistical approach offers a direct test of our hypotheses and is also consistent with prior studies, making our results directly comparable with the previous sinewave and point-light talker learning literature (Jesse & Bartoli, 2018; Rosenblum et al., 2002; Sheffert et al., 2002). To account for inflated familywise error as a result of the multiple t-tests we performed, critical p-values were Bonferroni corrected.

On each trial, participants could select one out of ten possible talkers, thus performance at chance was 10% (proportion of correct identifications = 0.10). A one-sample t-test of participant scores revealed that participants learned talkers' identities during the training phase at above chance levels (M=0.41, SD=0.10), $t(18)$=13.554, p<0.010, r=0.957. However, an analysis of variance (ANOVA) found that talker identification accuracy varied significantly across talkers, $F(9, 180)$=13.927, p<0.01, η^2_p =0.410 (see Table 2). Post hoc one-sample t-tests were performed to test whether each talker was identified at above-chance levels during training. These comparisons revealed that all of the ten talkers were identified significantly above chance, at a Bonferroni-corrected $\alpha = 0.005$ (see Table 2).

Test phase

A one-sample t-test revealed that participants' mean identification accuracy was above chance, (M=0.390, SD= 0.26), $t(18)$=10.243, p<0.01, r = 0.928. These accuracy results are comparable to what was reported by Sheffert et al. (2002; M = 0.44) in their test of sinewave talker learning, which also used a set of ten talkers. Again, identification accuracy varied across talkers, $F(9, 180)$= 10.111, p<0.05, η^2_p = 0.336. Post hoc t-tests for each talker revealed that all talkers were accurately identified above chance (0.10), at a Bonferroni-corrected α = 0.005 (see Table 3). These results show that perceivers could learn to identify talkers from our sinewave stimuli.

Table 2 Sinewave speech talker identification performance one-sample t-tests comparing subject identification accuracy against chance (.10) during the training phase of Experiment 1a

Talker	Mean	SD	t	p
F1	0.44	0.16	9.263	< 0.001
F2	0.68	0.24	10.534	< 0.001
F3	0.61	0.15	14.82	< 0.001
F4	0.44	0.23	6.444	< 0.001
M1	0.45	0.22	6.935	< 0.001
M2	0.23	0.09	6.296	< 0.001
M3	0.43	0.25	5.754	< 0.001
M4	0.34	0.13	8.047	< 0.001
M5	0.24	0.17	3.59	0.001
M6	0.22	0.11	4.755	< 0.001

Bonferroni-corrected alpha is p = .005.

Table 3 Sinewave speech talker identification performance one-sample t-tests comparing subject identification accuracy against chance (.10) during the test phase of Experiment 1a

Talker	Mean	SD	t	p
F1	0.26	0.14	4.982	< 0.001
F2	0.7	0.26	10.059	< 0.001
F3	0.5	0.22	7.925	< 0.001
F4	0.46	0.28	5.604	< 0.001
M1	0.57	0.26	7.88	< 0.001
M2	0.24	0.1	6.102	< 0.001
M3	0.35	0.32	3.405	0.0016
M4	0.26	0.16	4.359	< 0.001
M5	0.3	0.17	5.128	< 0.001
M6	0.26	0.13	5.365	< 0.001

Bonferroni-corrected alpha is $p = .005$

Experiment 1b: Point-light speech

Experiment 1b investigated whether participants could learn to identify novel talkers based on the isolated kinematic information in point-light displays. This experiment goes beyond the extant literature in which it has been found that point-light speech can support talker-identity learning for a small (four talkers) set of talkers (Jesse & Bartoli, 2018); here we tested talker learning with a larger (ten talkers) set of talkers.

To ensure that participants did not learn to identify point-light talkers using the idiosyncratic characteristics of the point-light videos that are unrelated to the talkers' kinematics, Rosenblum et al. (2007b; see also Jesse & Bartoli, 2018) included a condition in which participants were required to identify talkers from *static* point-light faces. Participants were unable to match static point-light faces to the talkers' identities, suggesting that important articulatory style information is preserved in dynamic time-varying kinematic information. We similarly employed a static point-light condition in the current investigation.

In addition, in order to ensure that participants were unable to identify talkers based on any idiosyncrasies in the positioning of point-lights, nine different point-light arrangements were implemented for each talker (e.g., Rosenblum et al., 2002; see also Fig. 2). Finally, to ensure that participants could not base their identification judgments on idiosyncrasies of the particular visual stimuli, they were tested on a different set of tokens (with different point-light configurations) from those on which they were trained.

Method

Experiment 1b was designed to closely match Experiment 1a. In place of sinewave speech, Experiment 1b used point-light speech. The point-light stimuli of Experiment 1b were derived from the same recordings used to make the sinewave speech

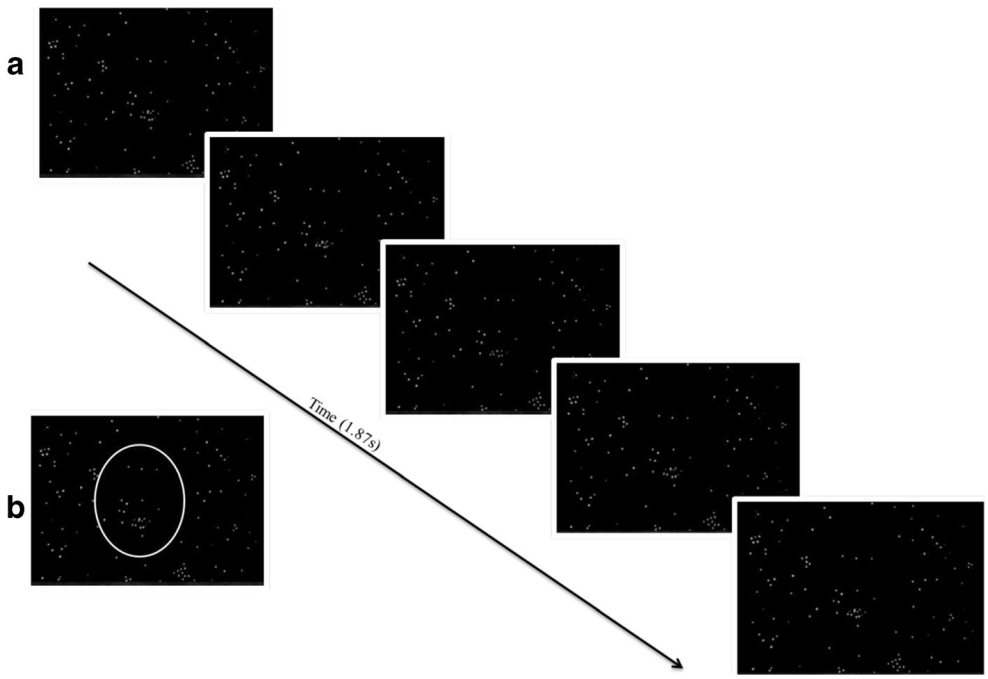

Figure 2 Panel A presents frames taken from the point-light speech of the talker F2 saying "The football game is over" to illustrate the dynamic point-light speech stimuli. Panel B indicates the approximate location of the talker's face within the point-light display. As can be seen in Panel A, these are the only points that move during the presentation of dynamic point-light speech (there are no moving dots during static point-light speech). The dots indicated in Panel B are also the dots whose configuration changed across talkers and across utterances within each talker

of Experiment 1a. While the procedures of this experiment closely match Experiment 1a, the use of point-light speech did result in subtle differences which are noted below. Furthermore, this experiment included two conditions: dynamic point-light speech and static point-light speech (see *Materials* section).

Participants Thirty-four undergraduates from the University of California, Riverside participated in this experiment, 18 in the dynamic point-light condition (ten female) and 16 in the static point-light condition (nine female). As with Experiment 1a, these sample sizes are consistent with past studies of talker-identity learning (e.g., Jesse & Bartoli, 2018; Sheffert et al., 2002). Participants received course credit for participation. All participants reported normal or corrected-to-normal hearing and vision. All participants were native speakers of North American English.

Materials The stimuli were derived from video components of the recordings that were used to generate the sinewave speech of Experiment 1a. General information on these recordings can be found above (see *Methods* section for Experiment 1a).

Dynamic point-light speech. Point-light stimuli were generated from the visual component of the same recordings that were used to make the sinewave speech of Experiment 1a. To create the point-light displays, the talkers were filmed against a black background with 30 fluorescent dots (cut out of construction paper covered in fluorescent paint) adhered to their faces, teeth, and tongue-tip (see Rosenblum et al., 2002, for additional details). To illuminate the points, two black-light (fluorescent) 24-in., 10-W bulbs were positioned vertically 3 ft away and at a 45° angle to the side/front of the face; no other lighting was used. The fluorescent dots were each 0.12-in. in diameter and were small enough so that they did not interfere with articulation. Dot placement was chosen based on a number of considerations (for details, see Rosenblum et al., 2002). First, locations were chosen that were known to convey good visual speech information based on previous research (e.g., Rosenblum & Saldana, 1996; Rosenblum, Johnson, & Saldana, 1996; Rosenblum et al., 2002). Fifteen dots were placed on the cheeks, forehead, and jawline of the talker's *face*. An additional 15 dots were placed on the talker's *mouth*, including teeth, tongue, and lips (see Rosenblum et al., 2002).

For the current study, effort was also made to hide any static/structural facial characteristics that might inform about talker identity (see Rosenblum et al., 2002, 2006, 2007b). Accordingly, for each utterance for each talker a *different configuration* of the 30 dots were placed within the face, tongue, teeth, and mouth areas. For each configuration, the positions created a quasi-random pattern so that talkers could not be easily identified by idiosyncratic facial dimensions (e.g., width). Additionally, during filming, each talker placed their face inside a cut-out hole located in the center of a black board with four plastic masks attached to it. The board and the masks were covered in the same fluorescent dots used on the talkers and created a static point-light background that contained the 3-D structure of multiple faces (Rosenblum et al., 2002, 2006, 2007b). The video image was composed of the entirety of this board with a talker's face in the middle (see also Fig. 2).

As noted above each talker was recorded nine times uttering the sentence and each time with a different quasi-random dot configuration (Rosenblum et al., 2002). This was done to prevent participants from memorizing a talker's dot configuration (see Fig. 3). Between each of these recordings, the dots were removed from the talker's face, then replaced elsewhere on the talker's face (again in a quasi-random configuration) and then filmed. During stimulus presentation, one of the dot configurations was presented in a familiarization phase, four different dot configurations were presented during a training phase, and four different configurations were presented during a test phase.

Nine video stimuli from nine talkers and eight video stimuli from one talker (89 total) were digitally captured on an Apple iMac for editing and presentation (one video for one of the talkers was of very low quality and thus was removed from the set of stimuli). Once digitized, video contrast was adjusted using Final Cut Pro software such that only the fluorescent dots, but not the talker's face, was visible. During the experiment, point-light speech for each talker was presented for the same duration as the corresponding sinewave speech version of the utterance from Experiment 1a.

Static point-light speech. Static frames were extracted from each of the 89 video clips using Final Cut Pro software. To ensure that the image contained no more than a minimal amount of articulatory information, a frame was chosen in which the talker had a static vowel and all points were visible, including those inside the mouth (e.g., Rosenblum et al., 2002, 2006, 2007b). During the experiment each static point-light image was presented for the same duration as the dynamic utterance from which it was extracted.

Procedure Like Experiment 1a, this experiment had three phases: familiarization, training, and testing. The general procedure of each of these phases closely followed that of Experiment 1a. Participants were first introduced to the point-light technique by being shown a static image taken from an unused point-light video. Participants were told that they would be asked to identify different talkers by associating point-light faces with talker names. The remaining procedures of the Familiarization, Training, and Test phases followed those of Experiment 1a. As for Experiment 1a, different sets of utterances were used for each talker in the Familiarization phase, Training phase, and Test phase. One utterance was presented during the Familiarization phase, four during the Training phase

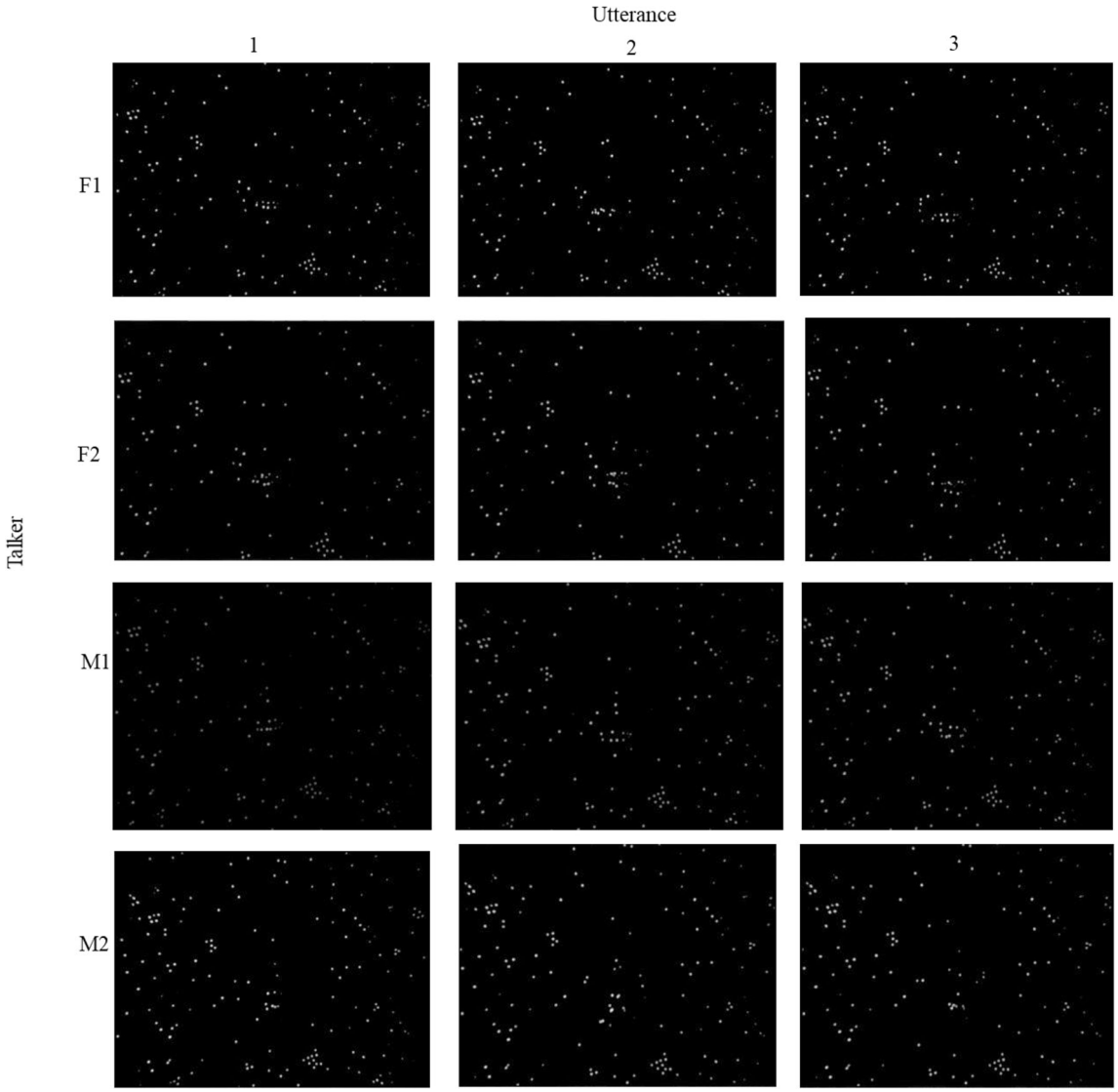

Fig. 3 Frames taken from the point-light speech of utterances of the sentence "The football game is over" from four (two male) of the talkers used in this study. Note the variability in point placement both between different talkers, but also within utterance from a single talker

(except for one talker for whom only three different utterances were used, one of which was repeated once for a total of four presentations), four during the Test phase.

Results

Training phase

A one-sample t-test revealed that participants in the *dynamic video condition* identified talkers at better than chance levels (M=0.350, SD=0.100), $t(17)$=10.972, p<0.05, r=0.936. Talker identification rates varied significantly across talkers, $F(9, 170)$=17.072, p<0.01, η^2_p=0.475. All but one talker was identified significantly above chance, and seven talkers were identified significantly above chance after correcting for inflated family-wise error using a Bonferroni correction at α=0.005 (see Table 4).

A one-sample t-test revealed that participants in the *static image condition* were also able to learn talkers at above chance levels, (M=0.19, SD= 0.06), $t(15)$=6.469, p< 0.01, r = 0.857 (Table 5). Here talker identification rates also significantly varied across talkers, F(9, 150)= 8.120, p<0.01, η^2_p =0.328, but Bonferroni-corrected t-tests revealed that only three out of ten talkers were identified significantly above chance, at Bonferroni-corrected α = 0.005 (see Table 5).

Test phase

The central question of this experiment is whether listeners could learn to identify talkers from point-light

Table 4 Dynamic point-light speech talker identification performance one-sample t-tests comparing subject identification accuracy against chance (.10) during the training phase of Experiment 1b

Talker	Mean	SD	t	p
F1	0.3	0.2	4.243	< 0.001
F2	0.24	0.11	5.400	< 0.001
F3	0.34	0.19	5.359	< 0.001
F4	0.24	0.21	2.828	0.0058
M1	0.22	0.18	2.828	0.0058
M2	0.49	0.28	5.909	< 0.001
M3	0.76	0.2	14.001	< 0.001
M4	0.24	0.13	4.569	< 0.001
M5	0.51	0.22	7.907	< 0.001
M6	0.15	0.13	1.632	0.0606

Bonferroni-corrected alpha is $p = .005$

Table 6 Dynamic point-light speech talker identification performance one-sample t-tests comparing subject identification accuracy against chance (.10) during the test phase of Experiment 1b

Talker	Mean	SD	T	p
F1	0.3	0.22	3.857	0.0006
F2	0.22	0.18	2.828	0.0058
F3	0.33	0.22	4.435	< 0.001
F4	0.23	0.22	2.507	0.0113
M1	0.17	0.13	2.284	0.0177
M2	0.45	0.28	5.303	< 0.001
M3	0.86	0.24	13.435	< 0.001
M4	0.19	0.12	3.182	0.0027
M5	0.55	0.28	6.819	< 0.001
M6	0.21	0.2	2.333	0.0161

Bonferroni-corrected alpha is $p = .005$

speech. A one-sample t-test revealed that participants in the dynamic face conditions could identify talkers at above chance levels, (M=0.35, SD=0.22), $t(17)=8.726$, $p<0.01$, $r = 0.904$. These results are comparable to what is reported by Jesse and Bartoli (2018; $M = .35$) in their test of point-light talker learning which included only four talkers (i.e., chance was .25). Talker-identification rates significantly varied across talkers, $F(9,170)=18.250$, $p<0.01$, $\eta^2_p = 0.491$ and Bonferroni-corrected t-tests ($\alpha = 0.005$) comparing mean identification of each talker to chance found that six of the ten talkers were identified above chance levels (see Table 6). It is unclear why some point-light talker identities were easier to learn than others, but talker differences have been observed in other point-light face recognition research (Jesse & Bartoli, 2018; Rosenblum et al., 2002, 2006, 2007b). Still, this overall pattern of results is consistent with those of Jesse and Bartoli (2018) in showing that point-light speech can support talker-identity learning.

For participants in the static face condition, a one-sample t-test revealed that overall mean talker identification accuracy was again above chance, (M=0.15, SD=0.05), t(15)= 4.0, p<0.01, $r = 0.718$. Mean talker identification accuracy was also calculated for each talker. Talker identification rates significantly varied across talkers, F(9, 150)= 4.732, p<0.01, $\eta^2_p = 0.221$; however, this difference is likely only driven by talker Male 2, who was identified at a much higher rate compared to the other talkers (M= 0.34, SD=0.23). Ten one-sample t-tests revealed that he was the only of ten talkers who was identified significantly above chance (Table 7). It is unclear why this one talker was identified so well from static stimuli.

Table 5 Static point-light speech talker identification performance one-sample t-tests comparing subject identification accuracy against chance (.10) during the training phase of Experiment 1b

Talker	Mean	SD	t	p
F1	0.18	0.12	2.667	0.0088
F2	0.18	0.11	2.909	0.0054
F3	0.18	0.08	4	0.0006
F4	0.13	0.09	1.333	0.1012
M1	0.21	0.19	2.316	0.0176
M2	0.43	0.2	6.6	< 0.001
M3	0.16	0.12	2	0.032
M4	0.18	0.09	3.556	0.0014
M5	0.15	0.07	2.857	0.006
M6	0.11	0.08	0.5	0.3122

Bonferroni-corrected alpha is $p = .005$

Table 7 Static point-light speech talker identification performance one-sample t-tests comparing subject identification accuracy against chance (.10) during the test phase of Experiment 1b

Talker	Mean	SD	t	p
F1	0.12	0.18	0.444	0.3315
F2	0.11	0.09	0.444	0.3315
F3	0.11	0.08	0.5	0.3122
F4	0.15	0.12	1.667	0.0582
M1	0.19	0.2	1.8	0.046
M2	0.34	0.23	4.174	< 0.001
M3	0.11	0.12	0.333	0.3717
M4	0.1	0.08	0	0.5
M5	0.15	0.07	2.857	0.006
M6	0.09	0.1	0.4	0.3474

Bonferroni-corrected alpha is $p = .005$

Importantly, faces were less accurately identified when they were presented as static point-light images compared to being presented as dynamic point-light videos. This was confirmed in a paired-samples t-test comparing average talker identification for each of our ten talkers, $t_{talker}(9) = 2.915$, $p = .009$, $r = .697$ (one-tailed; see also Fig. 4). Ten additional planned comparisons revealed that six out of ten talkers were identified at significantly ($p < .05$, uncorrected) higher rates in dynamic face stimuli compared to static (see Table 8). Collectively, these results indicate that talker identification benefited from the dynamic point-light displays. Additionally, this effect indicates that point-light speech is sufficient to support the learning of novel talkers.

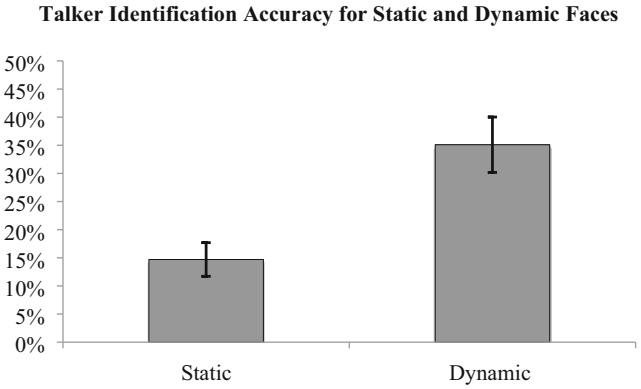

Fig. 4 Overall talker identification accuracy in dynamic face training group and static face training group (Experiment 1). Error bars indicate standard error of the mean

Experiment 1: Comparing visual and auditory unimodal results

The results of Experiment 1 suggest that participants can learn to identify talkers from sinewave replicas and point-light displays of speech utterances (see Tables 1, 2, 3, 4, 5 and 6). These experiments add to the evidence that dynamic time-varying talker information can be learned from both point-light and sinewave stimuli (see also Jesse & Bartoli, 2018; Sheffert et al., 2002). The current study expands on these prior studies in some important ways, notably using a larger point-light talker set than was used by Jesse and Bartoli (2018) and a shorter training period for sinewave speech talker learning than was used by Sheffert et al. (2002). Moreover, by using a single set of talkers and a comparable training paradigm for both Experiments 1a and 1b, we could compare talker-specific learning across point-light and sinewave speech conditions, something that was not possible in prior studies.

Talker-specific learning differences between sinewave and point-light speech were found. The results of ten planned comparisons, comparing performance on point-light and sinewave tests for each talker, show a significant difference between the point-light and sinewave tests for six of the ten talkers at p<0.005 levels (see Fig. 5). However, as illustrated in Fig. 5, these differences do not reflect a consistent *modality advantage* ($t_{talkers}[9] = 0.405$, $p=0.347$, $r = 0.134$ [one-tailed]). Some talkers were more accurately identified from their point-light displays while others were more accurately identified from their sinewave replicas.

Table 8 Test phase talker-identification accuracy differences comparing static to dynamic point-light speech

Talker	Difference	T	p
F1	-0.18	-2.606	0.013
F2	0.11	2.25	0.031
F3	0.22	3.868	<0.001
F4	0.08	1.314	0.198
M1	-0.02	-0.345	0.732
M2	0.11	1.249	0.22
M3	0.75	11.504	<0.001
M4	0.09	2.569	0.015
M5	0.4	5.704	<0.001
M6	0.12	2.209	0.034

This pattern of results suggests that there may be talker-specific information that is differentially salient in the auditory and visual modalities. In other words, these results indicate that talker-specific dimensions may determine whether identifications are more accurate for point-light or sinewave speech. Because both auditory and visual identification tests were conducted using a single set of talkers, we were able to perform additional analyses to examine whether common patterns of talker learning occurred across modalities. To examine this, we first computed the talker confusions of each talker (the rate at which a talker was identified correctly and the rate at which a talker was misidentified as each other alternative talker) in each unimodal condition (sinewave and dynamic point-light). We then examined the extent to which the talker confusions in sinewave speech related to the talker confusions in dynamic point-light displays for each talker. Finding a relationship between talker confusions in sinewave speech and dynamic point-light displays could indicate that perceivers use a common type of information to identify talkers in both modalities. If true, this common information across point-light and sinewave speech might support cross-modal learning of talker identification. Talker confusions are reported in Tables 9 and 10.

Pearson product-moment correlations of talker-confusions revealed a significant positive relationship between sinewave and dynamic point-light talker confusions for seven out of the ten talkers, and a significant average correlation (r_{ii} Spearman-Brown "Down") across talkers, $r = 0.578$, $p = 0.04$ (Rosenthal & Rosnow, 1991). These correlations suggest that talkers were often confused with the same other talkers across sinewave and point-light speech. This could suggest that common information

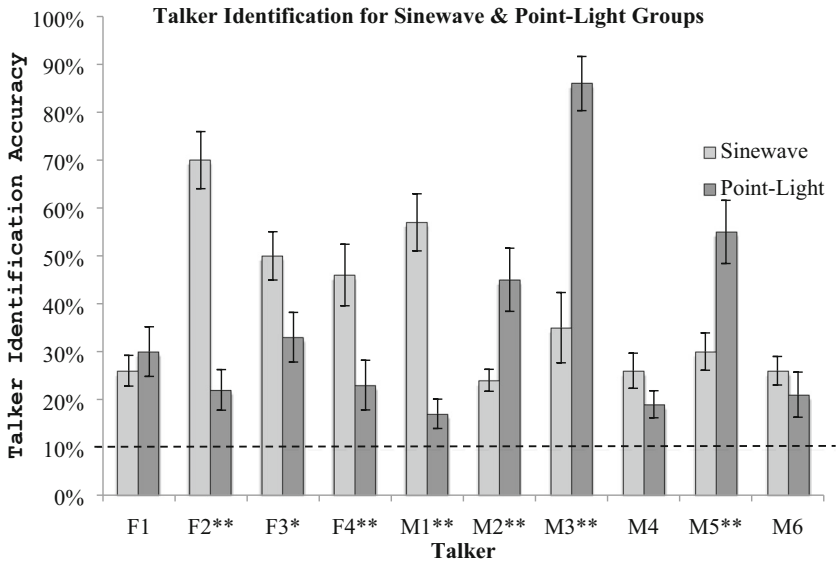

Fig. 5 Point-light and sinewave speech test performance by talker (Experiments 1). ** indicates p-values below the Bonferroni-corrected alpha of $p = .005$; * indicates $p < .05$, a marginally significant effect. Error bars indicate the standard error of the mean. The broken line indicates chance performance (10%)

was used in learning to identify talkers in both modalities, which could bode well for cross-modal talker learning. The possibility of common information across modalities is next examined in the context of talker-identity matching.

Experiment 2: Matching talkers across point-light and sinewave speech

The correlations between point-light and sinewave speech talker confusions observed in Experiment 1 suggest that perceivers can distinguish talkers from one another based, partly, on similar talker-specific characteristics shared across modalities. Experiment 2 used another method to test whether there is common, amodal talker information available across our point-light and sinewave speech stimuli: cross-modal talker matching.

Lachs and Pisoni (2004c) previously found that perceivers can match talkers' point-light speech with their sinewave speech, and vice versa, suggesting that enough common talker-specific information is available across point-light and sinewave speech to support cross-modal matching. To further

Table 9 The confusion matrix for sinewave speech. Each cell displays the probability that a given stimulus (the target) will be identified as a given talker identity (the response)

Sinewave Speech Confusion Matrix

		Response									
	Talker	F1	M1	M2	M3	F2	F3	M4	M5	F4	M6
Target	F1	**0.255**	0.011	0.013	0.016	0.032	0.363	0.021	0.011	0.250	0.021
	M1	0.000	**0.553**	0.118	0.092	0.011	0.000	0.076	0.029	0.016	0.087
	M2	0.005	0.124	**0.237**	0.216	0.016	0.000	0.095	0.097	0.021	0.168
	M3	0.003	0.205	0.126	**0.347**	0.000	0.003	0.095	0.071	0.003	0.137
	F2	0.047	0.011	0.013	0.008	**0.687**	0.061	0.005	0.011	0.121	0.024
	F3	0.158	0.013	0.024	0.024	0.047	**0.492**	0.024	0.034	0.155	0.021
	M4	0.016	0.061	0.129	0.116	0.008	0.026	**0.258**	0.234	0.018	0.121
	M5	0.016	0.068	0.105	0.121	0.011	0.016	0.224	**0.289**	0.042	0.095
	F4	0.161	0.011	0.029	0.018	0.158	0.071	0.026	0.024	**0.455**	0.039
	M6	0.016	0.061	0.203	0.166	0.016	0.013	0.161	0.076	0.018	**0.261**

The underlined values identify the probabilities of correct target-response matches, and the bolded responses identify the highest probability for a given response (within a column)

Table 10 The confusion matrix for point-light speech. Each cell displays the probability that a given stimulus (the target) will be identified as a given talker identity (the response)

Point-Light Speech Confusion Matrix											
		Response									
	Talker	F1	M1	M2	M3	F2	F3	M4	M5	F4	M6
Target	F1	**0.297**	0.072	0.036	0.022	0.067	0.114	0.108	0.089	0.103	0.089
	M1	0.047	**0.164**	0.067	0.225	0.097	0.072	0.086	0.022	0.119	0.089
	M2	0.053	0.067	**0.453**	0.089	0.053	0.053	0.067	0.022	0.053	0.083
	M3	0.003	0.014	0.028	**0.856**	0.008	0.028	0.011	0.006	0.014	0.022
	F2	0.086	0.111	0.047	0.019	**0.219**	0.117	0.125	0.025	0.089	0.147
	F3	0.036	0.094	0.069	0.022	0.133	**0.328**	0.094	0.014	0.108	0.097
	M4	0.103	0.128	0.039	0.017	0.103	0.106	**0.189**	0.017	0.167	0.128
	M5	0.086	0.033	0.061	0.014	0.033	0.039	0.078	**0.547**	0.053	0.053
	F4	0.086	0.125	0.061	0.050	0.064	0.089	0.047	0.031	**0.233**	**0.214**
	M6	0.094	0.086	0.067	0.011	0.089	0.106	0.097	0.094	0.142	0.208

The underlined values identify the probabilities of correct target-response matches, and the bolded responses identify the highest probability for a given response (within a column)

test whether shared idiolectic information exists across our own point-light and sinewave speech stimuli, we modified the cross-modal-matching paradigm employed by Lachs and Pisoni (2004c) with our own stimuli. Unlike the Lachs and Pisoni (2004c) study, every trial of the current experiment presented different utterances of the same sentence rather than the same exact utterance. That is, while Lachs and Pisoni (2004c) were able to show that participants can cross-modally match the same utterances, we tested if participants could match talkers across different utterances of the same sentence. Furthermore, Lachs and Pisoni (2004c) relied on point-light and sinewave speech stimuli taken from a set of four talkers, while our experiment included utterances from a set of ten talkers. Moreover, unlike Lachs and Pisoni (2004c), we used multiple point-light configurations for each talker. Collectively, these differences likely made our tests more challenging for participants. However, these changes also ensured that participants were not making cross-model matches based on the idiosyncratic characteristics of specific utterances (e.g., utterance length or point-light positions) and had to match talkers in the presence of a more diverse set of foil talkers. If with these added controls participants can match talkers' sinewave and point-light speech, then there must exist some talker-specific articulatory information shared across our own sinewave and point-light speech stimuli.

Method

Participants Consistent with Lachs and Pisoni (2004c), 41 undergraduates from the University of California, Riverside participated in this study (23 female). Participants received Psychology course credit for participation. All participants were native speakers of North American English and reported normal hearing and normal or corrected-to-normal vision.

Stimuli and procedure The point-light and sinewave speech stimuli employed in Experiment 1 were used in a cross-modal talker-matching task similar to the XAB two-alternative force-choice task used by Lachs and Pisoni (2004c). (Following Lachs & Pisoni, 2004c, as well as other cross-modal talker-matching tests, no familiarization phase was used in this experiment.) Participants were divided into two groups based on whether they were matching a single point-light talker to one of two sinewave talkers (21 participants), or whether they were matching a single sinewave talker to one of two point-light talkers (20 participants).

Participants were given detailed instructions before completing the cross-modal matching task. Participants matching a single point-light talker to one of two sinewave talkers were instructed that they would see a silent (dynamic) point-light speech display (X) followed by two sinewave speech utterances (A & B) with no visual stimulus. They were told that their task was to indicate which of the two sinewave utterances was the same talker as that seen in the point-light video. While neither the A or the B stimulus was taken from the same utterance as the X, one was a different utterance from the same talker. This was explained to each participant in an effort to discourage cross-modal matching based (erroneously) on superficial similarities. Participants were advised that the task would be difficult at first, but to try their best on every trial. Participants were also informed that all stimuli would be utterances of a single sentence; "The football game is over."

The same instructions were given to participants matching a sinewave utterance (X) to one of two point-light utterances (A & B), with the only change related to the switching of the compared modalities.

For participants matching a single point-light talker to one of two sinewave talkers, each trail began with an "X" on a computer screen for 1,000 ms followed by a point-light talker. After presentation of the point-light talker, an "A" was presented on the screen for 1,000 ms, followed by presentation of a sinewave talker. Then, a "B" was presented on the screen for 1,000 ms, followed by another sinewave talker.

For participants matching a single sinewave talker to one of two point-light talkers, each trial began with an "X" on a computer screen for 1,000 ms followed by a sinewave talker. After presentation of the sinewave talker, an "A" was presented on the screen for 1,000 ms, followed by presentation of a point-light talker. Then, a "B" was presented on the screen for 1,000 ms, followed by another point-light talker. At the end of each trial, participants were instructed to indicate which talker, A or B, matched talker X using a keyboard. Trials were separated by a 1,000-ms inter-trial interval.

The target talker matching X was randomly assigned to either A or B on each trial. The alternative talker was randomly selected from the remaining nine talkers, with all talkers equally represented. Participants matched each talker 18 times across the 180-trial cross-modal-matching task. The procedure was executed using PsyScope software (Cohen, MacWhinney, Flatt, & Provost, 1993). Visual stimuli were presented on a 24-in. ViewSonic VX2450 at 60 Hz and 1,920 × 1,080 resolution and auditory stimuli were presented through Sony MDR-V600 headphones at a comfortable listening level of 70 dB SPL.

Results and discussion

One-sample t-tests were used to evaluate the degree to which perceivers were able to match point-light and sinewave talkers. Across talkers, perceivers were able to match each point-light talker to the correct one of two sinewave talkers at levels significantly above chance (0.5), $M = 0.529$, $SE = 0.010$, $t(20) = 2.738$, $p = 0.013$, $r = 0.273$. Perceivers were also able to match the sinewave talkers to the correct one of two point-light talkers at levels significantly above chance, $M = 0.536$, $SE = 0.010$, $t(19) = 3.530$, $p = 0.002$, $r = 0.396$. A mixed-design analysis of variance (ANOVA) with talker as a within-groups factor and experimental group as a between-groups factor found no main effect of talker ($F[9, 351] = 0.836$, $p = 0.583$, $\eta_p^2 = 0.021$) nor experimental group ($F[1,39] = 0.453$, $p = 0.505$, $\eta_p^2 = 0.011$). The interaction of talker and experimental group was found to be significant ($F[9, 351] = 5.697$, $p < 0.001$, $\eta_p^2 = 0.127$), suggesting some variability in cross-modal matching among talkers depending on the experimental group. The results are reported in Fig. 6.

The results of Experiment 2 replicate the findings of Lachs and Pisoni (2004c). Perceivers were able to match talkers' point-light and sinewave speech with our stimuli. The magnitude of the effect is modest, but the degree to which our participants were able to accurately match talkers' point-light and sinewave speech is consistent with the values measured by Lachs and Pisoni (2004c; $M = 0.541$ both conditions). This is impressive for the current experiment when considering that, as compared to the experiment of Lachs and Pisoni (2004c): (a) our talker set was more than twice as large, (b) our participants needed to cross-modally match across different utterances of the audio and video speech material; and (c) our participants never saw the same point-light configuration more than once.

These results suggest that there is some amodal talker-specific information available across our point-light and sinewave stimuli to support cross-modal matching. The common information available across modalities may provide a cross-sensory benefit to learning talker identities, such that learning to identify a talker in one modality may facilitate learning to identify that same talker in another. In Experiment 3, we trained perceivers to identify talkers by their point-light speech and then examine the extent to which this point-light training enhanced their ability to learn to identify the same talkers by their sinewave speech.

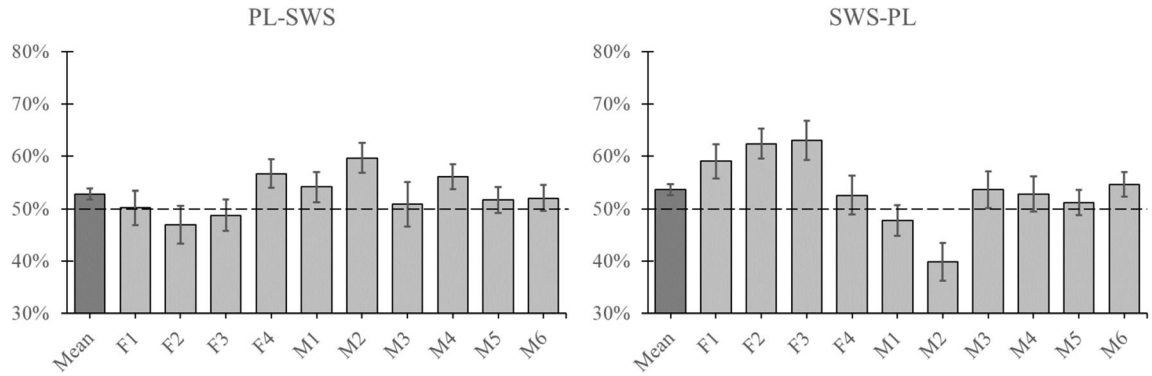

Fig. 6 Experiment 2 cross-modal-matching data. Error bars represent standard error of the mean

Experiment 3: Cross-modal talker-identity learning

The results of Experiment 1 indicated that the dynamic information contained in both point-light displays and sinewave speech supports talker-identity learning. The correlations between sinewave and point-light talker confusions further suggest that some of this information may be shared across sinewave and point-light speech. This point is re-enforced by the findings of Experiment 2 showing that participants were able to make explicit cross-modal talker matches. Experiment 3 directly tests the main question of our investigation: Can talker *learning* transfer across modalities?

Unlike Experiment 1, there was no test phase, as such, in this experiment. Instead, two training phases with feedback were used. Participants were trained to identify point-light faces during the first phase and were then trained again to identify sinewave voices in the second phase. The purpose of this design was to test if the training to identify specific talkers during the first phase facilitated training for identifying those same talkers in the second phase, despite each phase utilizing a different modality. Half of the participants were trained on the same set of talkers across the point-light and sinewave blocks and the other half of participants were trained on a different set of talkers across the point-light and sinewave blocks. If by the end of the sinewave training block talker identification was better for the "Same Talker" group, then experience with the point-light speech of the previous training block cross-modally facilitated the learning of the sinewave talkers.

Participants

Forty-nine undergraduates from the University of California Riverside participated in this study (24 in the "same-talker" condition; 27 females). Participants received Psychology course credit for participation. All participants reported normal hearing and normal or corrected-to-normal vision. All participants were native speakers of North American English.

Materials

The 89 dynamic point-light displays and the 89 sinewave utterances from Experiment 1 were used. Dynamic point-light faces were presented during the point-light training phase and sinewave voices were presented during the sinewave training phase. Stimuli were presented using PsyScope software (Cohen, MacWhinney, Flatt, & Provost, 1993). Other aspects of the stimulus presentation and experiment procedure were replicated from Experiment 1.

The point-light and sinewave training blocks were constructed in the following manner. First, in order to construct the two sinewave speech (Phase 2) blocks, identification results from Experiment 1a were used to construct two groups of five talkers (two women and three men) with nearly equal identification scores. (One group of five talkers was identified correctly 41% on average, while the other group of talkers was identified 42% on average). For each of these sinewave groups, two point-light (Phase 1) training blocks were constructed. One of these point-light blocks consisted of the same five talkers that were tested in the sinewave set, while the other point-light block consisted of the five talkers *not* tested in the sinewave block. This block arrangement allowed counterbalancing, such that all talkers: (a) could be tested in both point-light and sinewave forms; and b) could be presented in both same and different talker contexts across the two phases of the experiment.

Twenty-four subjects were randomly assigned to the "Same Talker" condition so that they learned the same point-light talkers in Phase 1 that they then learned as sinewave talkers in Phase 2. Twenty-five subjects were assigned to the "Different Talker" condition and learned a different set of point-light talkers in Phase 1 from those they later heard as sinewave talkers in Phase 2. If talker-identification learning can transfer across modalities, then subjects in the Same Talker condition should be able to better learn the group of sinewave talkers than those in the Different Talker condition.

Procedure

Initially, participants only received instructions for the point-light training phase and were not aware that they would later hear voices during the second part of the experiment. The experimenter briefly described the point-light technique, and participants were informed that they would not see normal faces during the experiment. Participants were given verbal instructions and were told that on-screen instructions would be displayed throughout the experiment. Only after completing the point-light training phase did participants learn that they would also be performing a sinewave speech-training task. The instructions given for this task followed the sinewave *training* instructions of Experiment 1a.

The 24 participants in the Same Talker condition were never told that they were presented with the same talkers across phases and different names were assigned to the talkers in the different modalities. The purpose of this was to minimize the possibility of participants recognizing superficial talker-learning cues, such as utterance duration, during the second training phase. The 25 participants in the different-talker condition were presented two different sets of talkers with two different sets of names.

Point-light phase Participants were first familiarized with five point-light talkers in the manner used for Experiment 1b. Participants were presented with two repetitions of two silent

point-light videos from five talkers for a total of 20 familiarization trials. Participants were then trained to identify five point-light talkers (see Experiment 1 training phase *Procedure* for details). Participants were presented with eight repetitions of four silent videos from each of the five talkers for a total of 160 trials with feedback.

Sinewave phase After the point-light training, participants were told they would be identifying talkers by listening to their voices. Participants were also told that the voices would not sound like normal voices and that vocal cues that are normally used to identify talkers were removed. The experimenter gave verbal instructions of the task and reminded the participant that on-screen instructions would also be displayed during the experiment.

The sinewave phase began with a set of familiarization trials. On each of these trials, participants were first presented with the name of the talker followed by a sinewave sentence. Participants were presented with two repetitions of two utterances of five voices for a total of 20 trials. Following this familiarization phase, participants were asked to identify the five talkers. For the sinewave training, participants were presented with eight repetitions of four sinewave utterances of five sinewave talkers for a total of 160 trials each with feedback. For the same-talker group, the four utterances were different from those used in the point-light phase.

Results

Point-light and sinewave talker identification For point-light training, participants in both the same-talker and different-talker groups identified the talkers at better than chance (20%): Different-talker group, M=0.49, SD=0.12, $t(24)$=11.827, p<0.001, r = 0.924; same-talker group, M=0.53, SD= 0.150, $t(23)$= 10.950, p< 0.001, r = 0.916. Most talkers were identified at better than chance for both groups, save for one talker (F1) in the same-talker group, M = .39, SD = .20, $t(11)$= 3.315, p = 0.007, p< 0.001, r = 0.707, corrected α = 0.005. The same-talker and different-talker groups showed no significant differences for identifying any individual talker during the point-light training phase (all p-values > 0.05). The similarity in the results of the same-talker and different-talker condition for point-light training is unsurprising as the distinction between these conditions is only implemented during the following sinewave speech training phase of the experiment.

For sinewave training, participants in both the same-talker and different-talker groups similarly identified talkers at better than chance (20%): Different-talker group, M=0.65, SD=0.16, $t(24)$=13.782, p<0.001, r = 0.942; same-talker group, M=0.70, SD= 0.11, $t(23)$= 22.523, p< 0.001, r = 0.978. All talkers were identified at above chance for both groups (corrected α = 0.005).

The results of Experiment 3 point-light and sinewave training replicate the results of Experiment 1, finding that participants could learn to accurately identify talkers from dynamic point-light displays and sinewave replicas of speech utterances.

Effects of point-light training on sinewave training To ensure that there were no talker-identification differences between the same-talker and different-talker participant groups prior to the critical sinewave speech training phase (where the distinction between these groups was manipulated), the point-light identification scores of each group were averaged for each talker and then compared. Talker-identification for point-light training was not found to depend on whether talkers were identified in the same-talker or different-talker conditions, $t(9)$=2.050, p=0.071, r = 0.564 (see also Fig. 7).

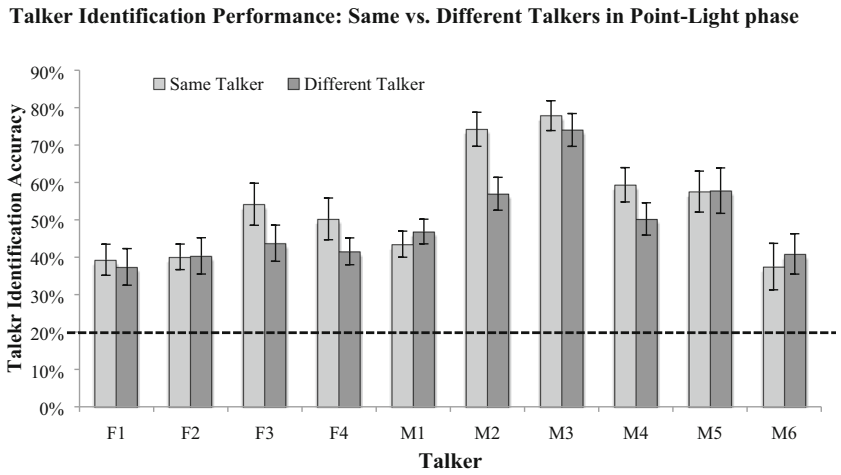

Fig. 7 Talker identification accuracy in the point-light speech phase of Experiment 3 as a function of when the talker was the "same" or "different" (note this distinction refers to a manipulation that occurred in a subsequent phase of the experiment). Error bars represent the standard error of the mean. The broken line indicates the 20% chance performance. All talkers in both groups were significantly above chance (all p-values < .0025), except for F1 (p=.007). Planned talker comparisons failed to find any single talker that was significantly better identified in the "Same Talker" group (all p-values >.005)

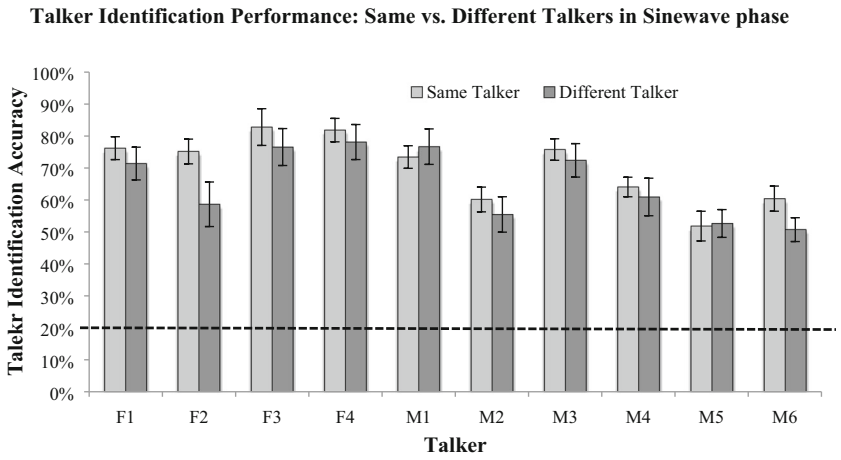

Fig. 8 Talker identification accuracy in the sinewave speech phase of Experiment 3 as a function of when the talker was the "same" as or "different" from a talker presented during the point-light phase. Error bars represent the standard error of the mean. The broken line indicates the 20% chance performance. All Talkers in both groups were significantly above chance (all p-values < .0025). Planned talker comparisons failed to find any single talker that was significantly better identified in the "Same Talker" group (all p-values >.005); it seems that there was a benefit of being in the "Same Talker" group that was small for each individual talker but consistent enough to produce a group-level effect

However, a similar analyses for sinewave training found that talkers *were* better identified when they had been trained in point-light (same-talker condition), $t(9) = 2.802$, p=0.020, r = 0.683 (see Fig. 8). While the overall size of the difference between the groups is small (~5%), it should be noted that it is a similar difference to what has been found for cross-modal talker-familiarity facilitation effects in other studies (i.e., Rosenblum et al., 2007a; Sanchez, et al., 2013). This effect is consistent with our hypothesis that learning to identify a talker through visible articulatory style can transfer across modalities to facilitate talker-identity learning through audible articulatory style.

General discussion

This study examined the use of articulatory information in cross-modal talker-identity learning. In Experiment 1, we tested whether observers could learn to identify unfamiliar talkers using isolated unimodal visible or audible articulatory information. Participants trained to identify novel talkers in sinewave speech or in dynamic point-light displays then performed a talker-identification test without feedback in the modality in which they were trained. Results from Experiment 1 suggest that the reduced talker information provided by sinewave and point-light speech could be used to learn talker identities. This experiment expanded on prior findings in showing talker learning in sinewave speech with a much shorter training period and using less language material than has been found before (e.g., Sheffert et al., 2002). Experiment 1 also showed talker learning with dynamic (and to a lesser degree, static) point-light speech with a much larger set of talkers than has been found previously (e.g., Jesse & Bartoli, 2018). Overall, these results suggest that talker-identity learning of these reduced speech stimuli is relatively robust.

Importantly, the design of Experiment 1 also allowed for a comparison of talker-learning patterns across point-light and sinewave speech. Using confusion matrices of unimodal test performance (Tables 9 and 10), this comparison showed that the pattern by which talkers are confused with specific "other talkers" is similar across point-light and sinewave speech. In other words, two talkers often confused in point-light speech were also likely to be confused in sinewave speech. This suggests that our participants may have been partially relying on a similar form of amodal information available across the modalities for their talker learning. If true, then cross-modal transfer of talker learning should be possible.

We further examined this possibility in Experiments 2 and 3. In Experiment 2 we found that in a two-alternative forced-choice (XAB) task, participants were able to correctly match talkers across modalities. This experiment expanded on the results of Lachs and Pisoni (2004) by showing cross-modal talker matching with sinewave and point-light speech in a larger set of talkers. Furthermore, that this match was made across different utterances indicates that it was based on information about the talker rather than idiosyncrasies of an individual utterance.

In Experiment 3 we investigated if this cross-modally available talker information was sufficient to support cross-modal talker learning from point-light to sinewave speech. One group of participants was trained to identify the same set of talkers across the modalities while the other group trained to identify two different sets of talkers across modalities. These data were analyzed as the average identification accuracy for talkers when they were familiar or novel to the participant. The talker data show a reliable, though small,

cross-modal effect of familiarity: Talker identification in the new modality was significantly better when the talkers were familiar than when they were novel. The results of Experiment 3 indicate that the information for talker identity was not only available in point-light speech, but that this information is in some way cross-modally available to facilitate talker learning of sinewave speech.

For Experiment 3, we chose to investigate the effect of visual-only training on the learning of auditory-only talker identification. This choice was made based on the sizable literature showing enhancement of auditory speech perception from visual information (e.g., Arnold & Hill, 2001; Grant & Seitz, 2000; Reisberg, McLean, & Goldfield, 1987; Sumby & Pollack, 1954; and see Rosenblum, in press, for a review). A reasonable question is whether training with auditory-only speech would likewise transfer to facilitate learning of visual-only speech. The bi-directionality of other cross-modal transfer effects has previously been established between studies (e.g., Rosenblum et al., 2007a; Sanchez et al., 2013), and this may be a reasonable approach for pursuing this interesting follow-up question.

Cross-modal information for talker identification

The most basic implication of these results is that perceivers are able to extract and learn cross-modal talker information from highly reduced auditory and visual displays. However, point-light and sinewave speech do not simply degrade the speech signal, but degrade it in a way that preserves information about the talker's articulations (e.g., Lachs & Pisoni, 2004; Remez, et al., 1997; Rosenblum, et al., 2007b). Sinewave speech eliminates features such as natural vocal timbre and fundamental frequency while retaining time-varying articulatory characteristics (Remez et al., 1981). Likewise, point-light speech removes the majority of the visual signal (facial features and configurations; face shape) but retains the patterns of facial motion during articulation (e.g., Rosenblum & Saldana, 1996). Both sinewave and point-light speech are known to inform about talker as well as speech perception. The current findings show that these stimuli can also provide amodal talker information allowing for cross-modal talker matching (Lachs and Pisoni, 2004) and transfer of talker learning.

The question arises of which specific talker dimensions may have been learned so as to allow for cross-modal transfer of talker learning. While the current research was not designed to address this question in detail, some speculation is warranted. One of the most conspicuous differences between talker's styles is in speaking rate. In the current research, speaking rate would be reflected in utterance duration/length because all talkers spoke the same sentence throughout. Clearly, utterance length is a dimension available in both point-light and sinewave stimuli. A small sampling of our talkers' utterance lengths can be seen Fig. 1 (see also Table 1). These examples show that while exact length varies from one talker's utterance to their next, it is likely that some talkers did have a generally faster speaking style than others. Potentially, subjects could use this dimension to help them learn to identify talkers in both modalities.

To examine whether subjects may have used speaking rate/utterance length in identifying talkers, analyses were conducted to determine if utterance lengths may have accounted for the response confusions measured in Experiment 1. For these purposes, we computed the average utterance-length difference between each pair of talkers, such that a smaller utterance-length difference indicated that those two talkers had more similar utterance lengths. Average Pearson product-moment correlation tests (r_{ii} Spearman-Brown Down) found that these length differences did not correlate with the pattern of talker confusions observed for either the sinewave speech (r=0.025, p =0.473) or the point light speech (r=0.083, p=0.410) responses. Regarding the cross-modal relationship of talker confusions, we found that controlling for the utterance length similarities between talkers (partial correlations) did not change the general relationship we observed (Experiment 1) between sinewave and point-light confusions for talkers, r=0.566, p=0.044. That is, a similar pattern of confusions was observed across sinewave and point-light speech conditions even after talker similarities in utterance length were partialed out. Together, these analyses suggest that utterance lengths/speaking rates did not play a major role in learning to identify talkers either within or across modalities.

Alternatively, it has been argued that both sinewave and point-light speech stimuli retain information for *talker-specific phonetic detail* (e.g., Remez et al., 1997; Rosenblum et al., 2002). Phonetic detail would involve a talker's idiosyncratic manner of producing segments that could be reflected in both types of stimuli. It has been suggested that the use of phonetic information for both speech and speaker perception may underlie the well-known contingencies of speech perception on talker familiarity (e.g., Remez et al., 1997; Rosenblum et al., 2002). Thus, the fact that talker familiarity can facilitate perception of noisy speech, word memory, and word priming may be based on a common use of phonetic detail information for both speech and talker recognition functions. If talker familiarity is based on experience with the talker-specific phonetic detail contained in sinewave and point-light speech, then the current results suggest that this information can take an amodal form such that this experience can be transferred across modalities.

Thus, it may be that the talker-specific phonetic detail contained in both point-light and sinewave speech allowed subjects to learn the talkers within and across modalities (e.g., Jesse & Bartoli, 2018; Sheffert et al., 2002). In fact, there is evidence that phonetic details can be salient for talker

identification and learning (e.g., Allen & Miller, 2004; and for a review, see Smith, 2015). For example, talkers are known to differ in exactly when they begin to coarticulate anticipatory lip-rounding for vowel production (e.g., Perkell & Matthies, 1992). Potentially, these talker differences could have appeared in the */u/* production of the word "football" for our talkers, and, if so, would likely be available in both point-light and sinewave stimuli. Relatedly, the degree of coarticulatory assimilation is known to be different between talkers (e.g., Amerman & Daniloff, 1977; Bladon & Al-Bamerni, 1976), a dimension that could also be available in our point-light and sine-wave samples. It is also known that talkers differ on how they mark the boundaries between syllables versus words, with some using different degrees of duration and others leniting consonants to varying degrees (Smith & Hawkins, 2012). These talker-specific dimensions may also be available in the sinewave and point-light versions of our stimuli. Future research can be designed to test which of these and other dimensions might be salient for talker learning from these stimuli.

Before finishing discussion of the supportive information, it is worth considering the results of Experiment 1b showing that some minor talker identification learning occurred for static point-light images. In theory, this could mean that the cross-modal effects of Experiments 2 and 3 could have been supported by learning static point-light information. While this possibility cannot be ruled out, it is worth noting the static image effect observed in Experiment 1b appears to have been driven by a single talker, M2. It could be that for this talker, some aspect of the static point configuration allowed for learning (despite the multiple random point placements used to disguise this information). On the other hand, the effects of Experiments 2 and 3 appear to be less dependent on the effects of any single talker. Furthermore, it is unclear how static point-light talker information could transfer to sinewave talker learning as shown in Experiment 3. Thus, we would argue that the weak talker learning observed with M2's static stimuli did not underlie our cross-modal learning effects. Instead, it is more likely that the dynamic talker-specific phonetic detail available in both sinewave and point-light speech supported cross-modal learning (Tables 9 and 10).

Crossmodal transfer of talker familiarity

The results of this investigation also expand on prior findings showing that talker familiarity can transfer across modalities to facilitate phonetic perception (e.g., Rosenblum et al., 2007a; Sanchez et al., 2013). The results here show that talker familiarity can also cross-modally facilitate talker identification. In this sense, the findings are consistent with "supramodal" theories that claim both speech and talker perception functions can use amodal articulatory information available across the senses (Rosenblum, Dorsi, & Dias, 2016; Rosenblum, 2005, 2008).

However, it must be acknowledged that not all published studies have found evidence for cross-modal transfer of talker information (e.g., van der Zande, et al., 2014a; and see also van der Zande, et al. 2014b). For example, van der Zande and his colleagues (2014a) tested talker influences on cross-modal word priming. It is well known that words recently heard are more easily perceived than novel words, and that this difference is enhanced if the same talker's voice is used for both presentations (e.g., Bradlow & Pisoni, 1999; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). Van der Zande and his colleagues (2014a) tested whether this talker enhancement of word priming would also work *cross-modally*. For this purpose, they first presented Dutch words, produced by one of two talkers, auditorily (and as clear/non-distorted speech) to subjects who were asked to type these words out. Subjects were then presented these same words, along with others, via visible articulations. These visible words were either presented using the same talker used in the first phase or a different talker. Subjects were asked to lipread these words to the best of their ability. Given the difficulty of silent lipreading, subject performance was evaluated for the correct number of visemes (visible phonemes) from the word responses. Van der Zande and his colleagues found that while subjects were able to better lipread the words that they heard in the first phase, there was no effect of same versus different talker on performance. The authors conclude that experience with talker information did not transfer across modalities to facilitate word priming.

These results contrast with those of Sanchez et al. (2013), who *did* find auditory to visual transfer of talker information in facilitating word priming. Moreover, Sanchez and her colleagues (2013) found that subjects were also better at lipreading new words from the same speaker they had (unknowingly) just heard. It is unclear why the van der Zande et al. (2014a) and Sanchez et al. (2013) projects provided different talker facilitation results. There were multiple difference between the projects including: duration of experience with the talker during the auditory phase (longer for Sanchez et al., 2013); number of talkers seen during the lipreading phase (more for Sanchez et al., 2013); and the lexical frequency of the words tested in both phases (more varied for Sanchez et al., 2013). The methodological differences underlying the discrepant results can be examined in future research.

However, it would not be surprising to find that cross-modal talker effects are less robust than unimodal talker effects. As we have argued elsewhere (e.g., Rosenblum et al., 2007a; Sanchez et al., 2013), it is likely that the informational dimensions available in one modality will not completely overlap with those in another modality. This fact would naturally limit the amount of perceptual learning that could be transferred across modalities relative to that transferable unimodally.

In sum, the present study provides evidence that information for learning talker identity is available cross-modally and can take the form of talker-specific phonetic details. These results add to other findings in speech perception (Rosenblum et al., 2007b; Sanchez et al., 2013; van der Zande et al., 2014a) and nonspeech perception (e.g., Kitagawa & Ichihara, 2002; Konkle et al., 2009; Levitan et al., 2015; Matsumiya, 2013) showing that perceptual experience in one modality can be transferred to another modality. Together, this research adds to the growing support that in some important ways, the perceptual brain is agnostic with regard to sensory modality (e.g., Ricciardi, Bonino, Pellegrini, & Pietrini, 2014; Rosenblum et al., 2007b).

Author Notes This research was supported by NSF grant 1632530 to LDR.

Dominique Simmons is now at Dimensional Mechanics, Bellevue, WA

Josh Dorsi is now at Penn State College of Medicine, Hershey, PA.

James W. Dias is now at the Medical University of South Carolina, Charleston, SC.

We confirm that the above article reports all included conditions and measurements, and that no data were excluded from the reported analyses. Sample sizes were selected based on previous literature.

References

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America, 115,* 3171. https://doi.org/10.1121/1.1701898

Amerman, J. D., & Daniloff, R. G. (1977). Aspects of lingual coarticulation. Journal of Phonetics, 5(2), 107-113.

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339–355. https://doi.org/10.1348/000712601162220

Bladon, R. A. W., & Al-Bamerni, A. (1976). Coarticulation resistance in English/l. Journal of Phonetics, 4(2), 137-150.

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice-and face-recognition areas. *Journal of Neuroscience*, *31*(36), 12906-12915.

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Bradlow, A.R., & Pisoni, D.B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and itemrelated factors. Journal of the Acoustical Society of America, 106, 2074–2085.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers*, *25(2)*, 257-271.

Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics*, *59*(6), 839–849. http://www.ncbi.nlm.nih.gov/pubmed/9270359

Grant, K. W., & Seitz, P. F. P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197–1208. https://doi.org/10.1121/1.422512

Jesse, A., & Bartoli, M. (2018). Learning to recognize unfamiliar talkers: Listeners rapidly form representations of facial dynamic signatures. *Cognition*, *176*(March 2017), 195–208. https://doi.org/10.1016/j.cognition.2018.03.018

Jesse, A., & Saba, P. (August, 2017). *Learning to recognize unfamiliar talkers from the word-level dynamics of visual speech*. Paper presented at the annual meeting of Audio-Visual Speech Perception, Stockholm, Sweden.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice': Matching identity across modality. *Current Biology*, *13*(19), 1709-1714.

Kitagawa, N., & Ichihara, S. (2002). Hearing visual motion in depth. Nature, 416(6877), 172-174.

Konkle, T., Wang, Q., Hayward, V., & Moore, C. I. (2009). Motion aftereffects transfer between touch and vision. *Current Biology*, *19*(9), 745–750. https://doi.org/10.1016/j.cub.2009.03.035

Lachs, L. & Pisoni, D. B. (2004a). Crossmodal source identification in speech perception. *Ecological Psychology, 16,* 159-187.

Lachs, L. & Pisoni, D. B. (2004b). Crossmodal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 378-296.

Lachs, L., & Pisoni, D. B. (2004c). Specification of cross-modal source information in isolated kinematic displays of speech. *Journal of Acoustical Society of America, 116,* 507-518.

Levitan, C.A, Ban, Y. H. A., Stiles, N. R. B., & Shimojo, S. (2015). Rate perception adapts across the senses: evidence for a unified timing mechanism. *Scientific Reports*, *5*(1), 8857. https://doi.org/10.1038/srep08857

Matlab version 7.10.0. Natick, Massachusetts: The MathWorks Inc., 2010

Matsumiya, K. (2013). Seeing a haptically explored face: Visual facial-expression aftereffect from haptic adaptation to a face. *Psychological Science*, *24*(10), 2088–2098. https://doi.org/10.1177/0956797613486981

Nygaard, L.C., & Pisoni, D.B. (1998). Talker-specific learning in speech perception. Perception & Psychophysics, 60, 355–376.

Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1994). Speech perception as a talker-contingent process. Psychological Science, 5, 42–46.

Perkell, J. S., & Matthies, M. L. (1992). Temporal measures of anticipatory labial coarticulation for the vowel/u: Within-and cross-subject variability. The Journal of the Acoustical Society of America, 91(5), 2911-2925.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A speechreading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). Erlbaum.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Speaker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance,* 23, 651-666.

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 40-61.

Remez RE, Rubin PE, Pisoni DB, Carrell TD (1981). Speech perception without traditional speech cues. Science; 212:947–950.

Ricciardi, E., Bonino, D., Pellegrini, S., & Pietrini, P. (2014). Mind the blind brain to understand the sighted one! Is there a supramodal cortical functional architecture? Neuroscience & Biobehavioral Reviews, 41, 64-77.

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

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*(6), 405–409. https://doi.org/10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D., Dorsi, J., & Dias, J. W. (2016). The Impact and Status of Carol Fowler's Supramodal Theory of Multisensory Speech Perception. *Ecological Psychology*, *28*(4), 262–294. https://doi.org/10.1080/10407413.2016.1230373

Rosenblum, L. D., Johnson, J. A., & Saldana, H. M. (1996). Point-light displays enhance comprehension of speech in noise. *Journal of Speech, Language, and Hearing Research, 39,* 1159-1170.

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007a). Lip-read me now, hear me later: Cross-modal transfer of speaker familiarity effects. *Psychological Science, 18(5),* 392-396.

Rosenblum, L. D., & Saldana, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology and Human Perception Performance, 22(2),* 318-331.

Rosenblum, L. D., Smith, N. M., Nichols, S. M., Hale, S., & Lee, J. (2006). Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception & Psychophysics, 68,* 84-93.

Rosenblum, L.D., Niehus, R.P. Smith, & N.M. (2007b). Look who's talking: Recognizing friends from visible articulation. *Perception, 36*, 157-159.

Rosenblum, L. D., Yakel, D. A., Baseer, N., Panchal, A., Nodarse, B. B., & Niehus, R. P. (2002). Visual speech information for face recognition. *Perception & Psychophysics, 64,* 220-229.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (Vol. 2). New York: McGraw-Hill.

Sanchez, K., Dias, J. W., & Rosenblum, L. D. (2013). Experience with a speaker can transfer across modalities to facilitate lipreading. *Attention, Perception, & Psychophysics, 75,* 1359-1365.

Schall, S., & von Kriegstein, K. (2014). Functional connectivity between face-movement and speech-intelligibility areas during auditory-only speech perception. *PloS One*, *9*(1), 1–11. https://doi.org/10.1371/journal.pone.0086325

Seitz, A., & Watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognitive Sciences*, *9*(7), 329–334. https://doi.org/10.1016/j.tics.2005.05.010

Sheffert, S. M., Pisoni, D. B, Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize speakers from natural, sine wave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance,* 28(6), 1447-1469.

Smith, R. (2015). Perception of speaker-specific phonetic detail. In: Fuchs, S., Pape, D., Petrone, C. & Perrier, P (Eds.), *Individual Differences in Speech Production and Perception* (pp. 11–38). Frankfurt a. M.: Peter Lang.

Smith, R., & Hawkins, S. (2012). Production and perception of speaker-specific phonetic detail at word boundaries. *Journal of Phonetics, 40*, 213-233.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215.

Van Der Zande, P., Jesse, A., & Cutler, A. (2014a). Hearing words helps seeing words: A cross-modal word repetition effect. *Speech Communication*, *59*, 31-43.

Van der Zande, P., Jesse, A., & Cutler, A. (2014b). Cross-speaker generalisation in two phoneme-level perceptual adaptation processes. *Journal of Phonetics*, *43*, 38-46

von Kriegstein, K., & Giraud, A. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*(10), 1809–1820. https://doi.org/10.1371/journal.pbio.0040326

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, *17*(3), 367–376. https://doi.org/10.1162/0898929053279577