Learning to recognize unfamiliar talkers from the word-level dynamics of visual speech

Alexandra Jesse[1], *Paul Saba*[1]

[1]Department of Psychological and Brain Sciences, University of Massachusetts Amherst, U.S.A.

ajesse@psych.umass.edu

Abstract

Familiar speakers can be recognized from seeing their idiosyncratic realization of visual speech [1], suggesting the long-term storage of facial dynamic signatures for speakers. Frameworks of face perception postulate that these facial dynamic signatures are only available for familiar speakers, but not for unfamiliar speakers. Our recent work has shown [2], however, that participants can rapidly learn to recognize unfamiliar speakers from the dynamic information contained in their visual speech when uttering sentences. While sentences can inform about the talker-specific realization of prosody, rate, and phonetic detail, words primarily provide information about variation in fine-phonetic detail, leading listeners to focus on different types of idiosyncrasies in these two types of materials when learning about auditory voices [3]. The present study tested whether representations of facial dynamic signatures can be formed from seeing the phonetic detail contained in words being uttered in isolation. Participants were trained to recognize two speakers from the dynamic information provided by point-light displays of isolated words. Feedback was given during training. At test, participants were tested on the point-light displays presented during training and on point-light displays of new words. Participants learned to recognize speakers from the word-level dynamics of visual speech independent of linguistic content.

Index Terms: audiovisual speech; talker recognition; learning

1. Introduction

Speakers vary in the way they produce the speech sounds of their native language, e.g., [4]-[6]. Listeners are sensitive to this variability across talkers in both auditory and visual speech, e.g., [7]-[9]. Yet some consistency in speech production can be found within each speaker and listeners adjust to the idiosyncrasies of a speaker in both modalities, e.g., [10]-[13]. The idiosyncratic realization of auditory and visual speech is however also informative about the identity of a familiar speaker [1], [14]. Recently, we have shown that listeners can learn to recognize unfamiliar speakers from seeing the facial dynamics they produce while uttering sentences [2]. In the present study, we extend this work by testing whether listeners can also learn to recognize unfamiliar speakers from seeing talker-specific phonetic detail in their production of single words.

The primary cue to speaker identity in auditory speech is voice quality, e.g., [15], [16]. To show that talkers can be recognized from their idiosyncratic phonetic realization of auditory speech, acoustic properties contributing to the percept of voice quality, such as the fundamental frequency of the speaker, e.g., [17], [18], have to be eliminated from the speech material while preserving phonetic variation. This goal can be achieved by creating sine-wave replicas of natural speech samples. Sine-wave speech consists of pure tones following the centroid frequencies and amplitude of the formants in the original speech sample. Sine-wave speech preserves sufficient spectrotemporal phonetic information for speech recognition [19]. Despite lacking the acoustic correlates of voice quality in sine-wave speech, voices of familiar speakers can be recognized from sine-wave speech [14] and voices of unfamiliar speakers can be learned [20]. These results provide strong evidence that the idiosyncratic realization of speech can be exploited as information about the identity of the speaker. Importantly, the indexical information isolated in sine-wave speech can also be accessed in natural speech, as learning of speakers transfers between these two speech types [14], [20] and the perceived similarity between unfamiliar voices does not change whether judged based on natural speech or sine-wave replicas [21]. Together, these results suggest that idiosyncrasies in the phonetic realization of speech are not just accessible when isolated in sine-wave speech, but also in the natural speech listeners encounter in their daily lives.

Point-light displays are the equivalent of sine-wave speech in the visual modality. Point-light displays of visual speech are created by placing dots on the face of a video-recorded speaker. The motion of these dots is tracked and used to animate a similar configuration of dots on a neutral background (i.e., a face is no longer visible). Point-light displays preserve biological motion but discard all invariant facial identity cues [22]. Point-light displays contain sufficient phonetic information for visual-only speech recognition and to elicit the audiovisual benefit [23], [24]. Point-light displays thus constitute a test case of whether or not visual idiosyncrasies in the realization of speech can provide speaker identity information. Indeed, participants can obtain sufficient speaker information from point-light displays samples of sentences to be able to match them to the same speaker's fully-illuminated talking face [25]. This result also shows that the identity information conveyed by point-light displays is also accessible in fully illuminated faces. Critically, invariant identity information from the face does not necessarily override the role of dynamic facial information in speaker recognition. When the motion of different speakers reading poems was used to animate the same avatar face, participants were able to identify which two of three samples came from the same speaker [26]. These findings suggest that perceivers extract, and hold temporarily in working memory, the identity information provided by speakers' visual speech [25], [27].

In addition, humans store facial dynamic signatures of talking in long-term memory for familiar speakers. Participants can recognize their friends from seeing point-light displays of them producing a sentence [1]. Neural and behavioral frameworks of face perception [28]-[31] have postulated that facial dynamic signatures, that also include information about

the realization of visual speech, are stored as representations separate from face representations of invariant properties. Furthermore, the consensus view seems to be that facial dynamic information only contributes to the recognition of familiar speakers, and only if viewing conditions are poor [32]-[35], but does not seem to contribute to learning to recognize unfamiliar speakers, e.g., [30], [36]. Studies examining whether seeing motion in fully illuminated faces aids the learning of unfamiliar faces have not reliably produced benefits [37]-[41]. These studies have, however, not tested whether dynamic signatures are stored for unfamiliar faces, but rather whether seeing motion helps forming invariant face representations.

Our recent work has challenged the status quo by demonstrating that mental representations of dynamic facial signatures of talking can be formed rapidly for unfamiliar speakers [2]. Presenting only point-light displays of talking faces uttering sentences during training and test, participants learned to recognize two speakers and four speakers from limited exposure. These point-light displays were normalized in configuration and size, eliminating all invariant cues to identity. Critically, the formed representations allowed participants to recognize these speakers also from new utterances. Our results thus demonstrate that participants can learn to recognize the identity of unfamiliar speakers from the motion they produce while talking, thereby establishing abstract identity representations that allow the recognition of these speakers independent of the linguistic content of their speech.

In the present study, we further tested participants' ability to learn to recognize unfamiliar speakers from the dynamics of their visual speech. Unlike in the previous study, point-light displays of two speakers uttering short isolated words, rather than sentences, were presented. Sentences provide listeners with longer samples of talkers than isolated words. Longer samples allow for better learning of auditory voices [42], [43]. In addition to length, sentences and isolated words also differ in the types of idiosyncrasies they can inform about [3]. Sentences provide listeners with information about the talker-specific phonetic realization of individual speech sounds and words, but also with information about more global idiosyncrasies in the realization of prosody and speaking rate. Learning to recognize speakers from sentences therefore does not require attending to the talker-specific phonetic detail. When learning to recognize speakers from words, however, listeners must consider talker differences in the realization of fine-phonetic detail. Learning to recognize speakers from words should therefore be more difficult than learning to recognize speakers from sentences. Listeners also focus on different types of idiosyncrasies in sentences vs. words. While learning to recognize speakers by their auditory voices from both types of materials can provide a benefit for speech perception, this benefit is largest when listeners are tested on the same type of material as they had been trained on; i.e., isolated words or sentences [3]. In our prior study using sentences, participants may have therefore learned to recognize speakers from visual global idiosyncrasies, such as in the realization of prosody or speaking rate. In the present study, we tested whether participants could also learn to recognize speakers from their idiosyncrasies in the phonetic realization of speech sounds and words.

2. Experiment

2.1. Participants

Twenty-four monolingual native speakers of American English (five men; mean age = 20.25 years) with no reported language or attention deficit participated. All had normal hearing and (corrected-to-) normal vision.

2.2. Materials

Two sets of ten monosyllabic consonant-vowel-consonant English words each were created. Words in both sets contained a similar variety of visemes. Each word consisted of a unique viseme combination. Sets were matched on their average word frequency (M = 93.71, M = 89.94; $t(18) = 0.1$, $p = .92$)[44].

Twenty-three 3-mm dots of white construction paper were attached to the face of a male and a female native speaker of American English. Speakers' faces were illuminated using a mixture of ultraviolet and halogen lights [45]. Videos of the speakers producing the selected words in isolation were recorded as h.264 at 25 fps with a SONY EVI-HD7V camera. Audio was recorded at the same time in mono at a 48 kHz sampling rate, using a Shure KSM44A microphone. To create the point-light displays, the motion of the dots was tracked in Adobe After Effects CS5 and verified frame-by-frame in a visual check. The obtained motion paths were then used to animate an average dot configuration (see Figure 1), created by calculating the mean locations of the dots in the first frame of a selected video for each speaker. By averaging the dot configurations, any differences in size or shape of the faces and in the placement of the dots were eliminated, see also [27], [46].

Figure 1: *Average point-light display configuration.*

2.3. Procedure

Each participant was tested individually in a sound-attenuated booth. During an initial training phase, participants received point-light displays of words from one of the two sets. The same set was selected for both speakers. On each trial (see Figure 2), participants saw one point-light display before choosing by button press one of two displayed names (*Anna*, *Owen*). No sound was presented. Once participants had answered, their response was shown along with the correct name. If participants had responded incorrectly, they next had to answer once more with the correct name. Independent of accuracy, participants were then always shown the same point-light display again, along with the name of the speaker printed underneath. No response was collected for this second presentation. The amount of exposure was therefore the same for all participants. Each participant received three blocks, each consisting of a randomized presentation of all ten words from each speaker in a set. In total, each participant saw 120 point-light displays (i.e., 3 blocks x 10 words x 2 speakers x 2 presentations per trial).

In the subsequent test phase, participants were presented with the same twenty point-light displays they had studied during training (*familiar word condition*) as well as with twenty point-light displays of new words (*new word condition*) spoken by the same speakers. Presentation order was completely randomized. On each trial, participants saw one point-light display before choosing the name of the speaker from two options. No sound was presented. No feedback was given. Assignment of sets to training, and thus to condition at test, was counterbalanced across participants.

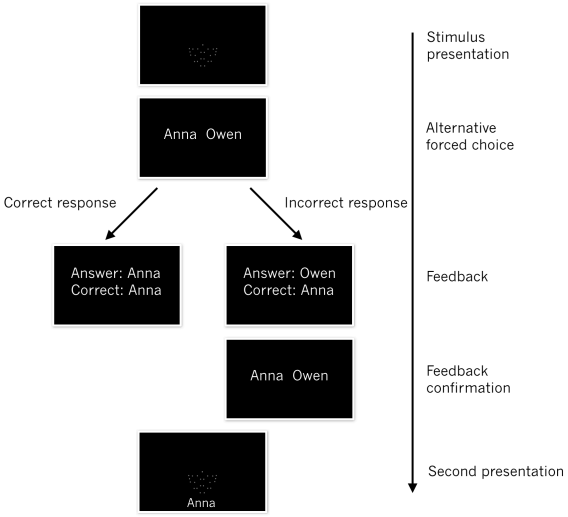

Figure 2: *Schematic representation of a training trial. Test trials only consisted of the stimulus presentation followed by the alternative forced choice.*

2.4. Results

2.4.1. Training

Figure 3 shows a histogram of participants' accuracy scores by training block. Only four participants were below chance level performance (.5) in the third block. All of them had performed above or at chance level on a previous block. A one-sample t-test comparing accuracy in the third block to chance showed that participants had learned to recognize the speakers from their facial dynamic signatures by the end of training (M = .63, SD = 0.13, $t(23)$ = 4.93, p < .0001; D = 1.01).

We further examined the build-up of learning: While participants did not reliably recognize speakers during the first block (M = .5, SD = 0.12, $t(23)$ = 0.17, p = .87; D = 0.03), learning became evident in the second block (M = .61, SD = 0.13, $t(23)$ = 4.41, p < .0002; D = 0.9). Paired two-sample t-tests comparing the change in performance across blocks showed that learning improved between the first two blocks ($t(23)$ = 3.6, p < .01; D = 0.88), but then remained at a similar level for the remainder of the training phase ($t(23)$ = -0.55, p = .58; D = 0.11). Participants had thus learned to recognize speakers within the first two blocks of exposure.

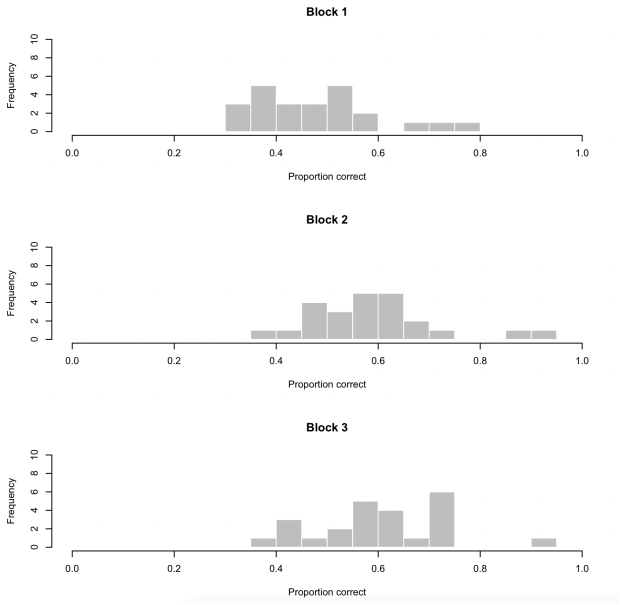

Figure 3: *Histogram of participants' accuracy scores across training blocks.*

2.4.2. Test

Figure 4 shows a histogram of participants' accuracy scores for recognizing speakers from point-light displays repeated from training and from point-light displays of new words. One-sample t-tests to chance level performance (.5) revealed that at test, participants were able to recognize the speakers from their facial dynamics when speakers were uttering the words already encountered during training (M = .61, SD = 0.13, $t(23)$ = 4.17, p < .001; D = .85) and when the words were new (M = .64, SD = 0.15, $t(23)$ = 4.51, p < .001; D = 0.92). Participants fully generalized their knowledge about the facial dynamics of a speaker to new materials, since there was no difference in recognition accuracy as a function of whether the point-light displays contained familiar or new words ($t(23)$ = -0.85, p = .4; D = 0.16).

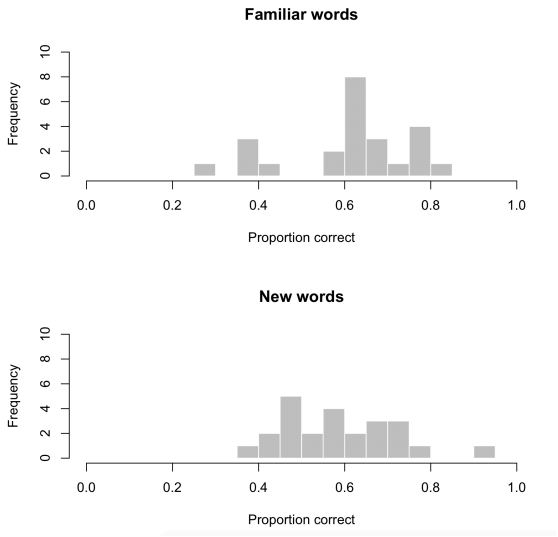

Figure 4: *Histogram of participants' accuracy scores by test condition.*

3. Discussion

Speakers vary in their realization of visual speech. The dynamics of visual speech provide identity information that can be learned [2] and stored in long-term memory [1]. In the present study, we have replicated our previous results that seeing the biological motion associated with the production of speech is sufficient for listeners to learn to recognize unfamiliar speakers based on their facial dynamic signatures. We extended our prior work by showing that not only sentences but also individual words provide the critical information needed to acquire abstract representations of facial dynamic signatures for unfamiliar speakers, that allow the recognition of speakers independent of the linguistic content of their speech. Together these findings challenge the assumption that representations of dynamic facial signatures, entailing information on how speakers talk, only play a role in the recognition of familiar speakers, e.g., [30], [36]. Instead, our results show that these representations become readily available from limited exposure for unfamiliar speakers, thus are in place to aid recognition.

The results of the present study critically extend our prior findings by demonstrating that participants can learn to recognize unfamiliar speakers from dynamic information isolated in the point-light displays of individual word productions. Learning to recognize speakers from spoken words rather than from sentences constitutes a more difficult test case, because words provide shorter speech samples of a speaker than sentences. Learning of auditory voices, for example, benefits from longer samples [42], [43]. In addition, sentences also provide information about global speaker attributes related to the idiosyncratic realization of prosody and speaking rate. In contrast, seeing words uttered instead of sentences forces participants to focus on fine-phonetic detail in the production of speech sounds and words [3]. Our results thus demonstrate that seeing the dynamics of spoken words provides sufficient talker-specific phonetic detail to create an identity representation of the speaker. This finding dovetails nicely with prior work showing that the identity information that can be extracted from spoken words is sufficient to match point-light displays and sine-wave speech samples of the same speaker [47]. Importantly, our study provides, in addition, evidence that listeners store the talker information obtained from the visual speech samples in long-term memory, and that the representations formed based on this information allow the future recognition of the speaker from their visual speech independent of its linguistic content.

Our results also support our prior finding that learning of facial dynamics signatures of talking can occur with limited exposure: Participants learned to recognize speakers within the second block of exposure to the training set. That is, participants as a group recognized speakers by the third presentation of a word, that is by the second trial per word, as each trial contained two presentations. Learning to recognize speakers from words was, however, more difficult than from sentences. In our prior work on sentences, learning was already completed within the first block of exposure (each trial consisted two presentations of a point-light display per trial). It is important to note, however, that the two studies differed in other aspects as well (e.g., speakers), and thus do not warrant a direct comparison. Exposure was more variable in the present study using words than in the previous study on sentences: In Jesse & Bartoli [2], participants received four tokens of two sentences each from each speaker during training. Test materials consisted of four different sentences, that is, eight tokens per sentence per speaker. In contrast, we presented participants in the present study with one token for each of ten different words per speaker during training and with 20 words at test. Exposure was therefore more variable in terms of linguistic content and less variable in terms of samples provided for each item. While repetitions of samples from a to-be-learned category aids learning to recognize the category from these samples, variability of samples leads to more robust generalization to other samples of the same category, e.g., [48]-[51], as it allows for better abstraction of the information shared across samples that indicates membership. While the contribution of variability to acquiring abstract speaker representations that allow reliable recognition of speakers independent of the linguistic content of their speech deserves further investigation, our two studies together demonstrate that representations of facial dynamic signatures can be acquired for unfamiliar speakers from very little exposure to word-level and sentence-level dynamics.

4. Conclusions

Speakers show systematic idiosyncrasies in their production of auditory and visual speech that provide identity information [1], [2], [14]. Listeners need very little exposure to speakers' visual speech to extract sufficient information to establish identity representations of the speakers' dynamic facial signatures. These representations allow recognition of speakers' identity from their visual speech, independent of the linguistic content of their speech. Listeners can establish these representations from seeing idiosyncrasies in the global attributes of spoken sentences, but also from talker-specific phonetic detail in the production of words.

5. Acknowledgements

This work was part of an undergraduate thesis conducted by the second author under the supervision of the first author.

6. References

[1] L. D. Rosenblum, R. P. Niehus, and N. M. Smith, "Look who's talking: recognizing friends from visible articulation," *Perception*, vol. 36, no. 1, pp. 157–159, 2007.

[2] A. Jesse and M. Bartoli, "Learning to recognize unfamiliar talkers: Listeners rapidly form representations of facial dynamic signatures," submitted.

[3] L. C. Nygaard and D. B. Pisoni, "Talker-specific learning in speech perception," *Percept Psychophys*, vol. 60, no. 3, pp. 355–376, 1998.

[4] J. S. Allen, J. L. Miller, and D. DeSteno, "Individual talker differences in voice-onset-time," *J Acoust Soc Am*, vol. 113, no. 1, pp. 544–9, 2003.

[5] R. S. Newman, S. A. Clouse, and J. L. Burnham, "The perceptual consequences of within-talker variability in fricative production," *J Acoust Soc Am*, vol. 109, no. 3, pp. 1181–1196, 2001.

[6] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J Acoust Soc Am*, vol. 24, no. 2, pp. 175–184, 1952.

[7] J. S. Allen and J. L. Miller, "Listener sensitivity to individual talker differences in voice-onset-time," *J Acoust Soc Am*, vol. 115, no. 6, pp. 3171–3183, 2004.

[8] S. L. M. Heald and H. C. Nusbaum, "Talker variability in audio-visual speech perception.," *Front Psychol*, vol. 5, no. 698, 2014.

[9] D. A. Yakel, L. D. Rosenblum, and M. A. Fortier, "Effects of talker variability on speechreading," *Percept Psychophys*, vol. 62, no. 7, pp. 1405–1412, 2000.

[10] P. Bertelson, J. Vroomen, and B. de Gelder, "Visual recalibration of auditory speech identification: a McGurk aftereffect.," *Psychol*

Sci, vol. 14, no. 6, pp. 592–597, 2003.
[11] D. Norris, J. M. McQueen, and A. Cutler, "Perceptual learning in speech.," *Cognitive Psychol*, vol. 47, no. 2, pp. 204–238, 2003.
[12] M. Baart and J. Vroomen, "Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading," *Neuroscience Letters*, vol. 471, no. 2, pp. 100–103, 2010.
[13] P. van der Zande, A. Jesse, and A. Cutler, "Lexically guided retuning of visual phonetic categories," *J Acoust Soc Am*, vol. 134, no. 1, pp. 562–571, 2013.
[14] R. E. Remez, J. M. Fellowes, and P. E. Rubin, "Talker identification based on phonetic information.," *J Exp Psychol Hum Percept Perform*, vol. 23, no. 3, pp. 651–666, 1997.
[15] J. Kreiman, D. Van Lancker-Sidtis, and B. R. Gerratt, "Perception of voice quality," in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds. John Wiley & Sons, 2008, pp. 338–362.
[16] J. Laver, *The phonetic description of voice quality*. Cambridge Studies in Linguistics London, 1980.
[17] R. Brown, "An experimental study of the relative importance of acoustic parameters for auditory speaker recognition.," *Lang Speech*, vol. 24, no. 4, pp. 295–310, 1981.
[18] C. LaRiviere, "Contributions of fundamental frequency and formant frequencies to speaker identification," *Phonetica*, vol. 31, no. 3, pp. 185–197, 1975.
[19] R. E. Remez, P. E. Rubin, D. B. Pisoni, and T. D. Carrell, "Speech perception without traditional speech cues.," *Science*, vol. 212, no. 4497, pp. 947–949, 1981.
[20] S. M. Sheffert, D. B. Pisoni, J. M. Fellowes, and R. E. Remez, "Learning to recognize talkers from natural, sinewave, and reversed speech samples.," *J Exp Psychol Hum Percept Perform*, vol. 28, no. 6, pp. 1447–1469, 2002.
[21] R. E. Remez, J. M. Fellowes, and D. S. Nagel, "On the perception of similarity among talkers," *J Acoust Soc Am*, vol. 122, no. 6, pp. 3688–3696, 2007.
[22] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Percept Psychophys*, vol. 14, no. 2, pp. 201–211, 1973.
[23] L. D. Rosenblum and H. M. Saldana, "An audiovisual test of kinematic primitives for visual speech perception.," *J Exp Psychol Hum Percept Perform*, vol. 22, no. 2, pp. 318–331, 1996.
[24] L. D. Rosenblum, J. A. Johnson, and H. M. Saldana, "Point-light facial displays enhance comprehension of speech in noise," *J Speech Lang Hear Res*, vol. 39, no. 6, p. 1159-1170, 1996.
[25] L. D. Rosenblum, D. A. Yakel, N. Baseer, and A. Panchal, "Visual speech information for face recognition," *Percept Psychophys*, vol. 7667, no. 57, pp. 479–487, 2002.
[26] C. Girges, J. Spencer, and J. O'Brien, "Categorizing identity from facial motion," *Q J Exp Psychol*, vol. 68, no. 9, pp. 1832–1843, 2015.
[27] R. J. Bennetts, J. Kim, D. Burke, K. R. Brooks, S. Lucey, J. Saragih, and R. A. Robbins, "The movement advantage in famous and unfamiliar faces: a comparison of point-light displays and shape-normalised avatar stimuli," *Perception*, vol. 42, no. 9, pp. 950–970, 2013.
[28] M. Bernstein and G. Yovel, "Two neural pathways of face processing: A critical evaluation of current models," *Neuroscience & Biobehavioral Reviews*, vol. 55, pp. 536–546, 2015.
[29] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini, "The distributed human neural system for face perception," *Trends in Cognitive Sciences*, vol. 4, no. 6, pp. 223–233, 2000.
[30] A. J. O'Toole, D. A. Roark, and H. Abdi, "Recognizing moving faces: a psychological and neural synthesis.," *Trends in Cognitive Sciences*, vol. 6, no. 6, pp. 261–266, 2002.
[31] V. Bruce and A. W. Young, "Understanding face recognition," *Br J Psychol*, vol. 77, no. 3, pp. 305–327, 1986.
[32] B. Knight and A. Johnston, "The role of movement in face recognition," *Visual Cognition*, vol. 4, no. 3, pp. 265–273, 1997.
[33] K. Lander and V. Bruce, "Recognizing famous faces: Exploring the benefits of facial motion," *Ecological Psychology*, vol. 12, no. 4, pp. 259–272, 2000.
[34] K. Lander, V. Bruce, and H. Hill, "Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces," *Appl Cognitive Psychol*, vol. 15, no. 1, pp. 101–116, 2001.
[35] K. Lander and V. Bruce, "Repetition priming from moving faces," *Memory & Cognition*, vol. 32, no. 4, pp. 640–647, 2004.
[36] V. Natu and A. J. O'Toole, "The neural processing of familiar and unfamiliar faces: a review and synopsis.," *Br J Psychol*, vol. 102, no. 4, pp. 726–747, 2011.
[37] F. Christie and V. Bruce, "The role of dynamic information in the recognition of unfamiliar faces," *Memory & Cognition*, vol. 26, no. 4, pp. 780–790, 1998.
[38] K. Lander and V. Bruce, "The role of motion in learning new faces," *Visual Cognition*, vol. 10, no. 8, pp. 897–912, Nov. 2003.
[39] G. E. Pike, R. I. Kemp, N. A. Towell, and K. C. Phillips, "Recognizing moving faces: the relative contribution of motion and perspective view information," *Visual Cognition*, vol. 4, no. 4, pp. 409–438, 1997.
[40] K. S. Pilz, I. M. Thornton, and H. H. Bülthoff, "A search advantage for faces learned in motion," *Exp Brain Res*, vol. 171, no. 4, pp. 436–447, 2005.
[41] K. S. Pilz, H. H. Bülthoff, and Q. C. Vuong, "Learning influences the encoding of static and dynamic faces and their recognition across different spatial frequencies," *Visual Cognition*, vol. 17, no. 5, pp. 716–735, 2009.
[42] G. E. Legge, C. Grosmann, and C. M. Pieper, "Learning unfamiliar voices.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 10, no. 2, pp. 298–303, 1984.
[43] T. L. Orchard and A. D. Yarmey, "The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification," *Appl Cognitive Psychol*, vol. 9, no. 3, pp. 249–260, 1995.
[44] M. Brysbaert and B. New, "Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," *Behav Res Methods*, vol. 41, no. 4, pp. 977–990, 2009.
[45] S. M. Thomas and T. R. Jordan, "Techniques for the production of point-light and fully illuminated video displays from identical recordings.," *Behav Res Methods Instrum Comput*, vol. 33, no. 1, pp. 59–64, 2001.
[46] H. Hill, Y. Jinno, and A. Johnston, "Comparing solid-body with point-light animations.," *Perception*, vol. 32, no. 5, pp. 561–566, 2003.
[47] L. Lachs and D. B. Pisoni, "Specification of cross-modal source information in isolated kinematic displays of speech," *J Acoust Soc Am*, vol. 116, no. 1, pp. 507–518, 2004.
[48] C. N. Wahlheim, B. Finn, and L. L. Jacoby, "Metacognitive judgments of repetition and variability effects in natural concept learning: evidence for variability neglect.," *Memory & Cognition*, vol. 40, no. 5, pp. 703–716, 2012.
[49] S. E. Lively, J. S. Logan, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories.," *J Acoust Soc Am*, vol. 94, no. 3, pp. 1242–1255, 1993.
[50] D. Homa, J. Cross, D. Cornell, D. Goldman, and S. Shwartz, "Prototype abstraction and classification of new instances as a function of number of instances defining the prototype," *J Exp Psychol*, vol. 101, no. 1, pp. 116–122, 1973.
[51] W. F. Dukes and W. Bevan, "Stimulus variation and repetition in the acquisition of naming responses.," *J Exp Psychol*, vol. 74, no. 2, pp. 178–181, 1967.