

**Running head:** Speaker identity supports phonetic category learning

**Speaker identity supports phonetic category learning**

Nivedita Mani & Signe Schneider

Language Acquisition Junior Research Group

University of Göttingen

Article in press: *Journal of Experimental Psychology: Human Perception and Performance*

**ABSTRACT**

Visual cues from the speaker's face, such as the discriminable mouth movements used to produce speech sounds, improve discrimination of these sounds by adults. The speaker's face, however, provides more information than just the mouth movements used to produce speech – it also provides a visual indexical cue of the identity of the speaker. The current paper examines the extent to which there is separable encoding of speaker identity in speech processing and asks whether speech discrimination is influenced by speaker identity. Does consistent pairing of different speakers' faces with different sounds i.e., hearing one speaker saying one sound and a second speaker saying the second sound, influence the brain's discrimination of the sounds? ERP data from participants previously exposed to consistent speaker-sound pairing indicated improved detection of the phoneme change relative to participants previously exposed to inconsistent speaker-sound pairing i.e. hearing both speakers say both sounds. The results strongly suggest an influence of visual speaker identity in speech processing.

Keywords: Phoneme discrimination, Faces, Audio-visual speech perception, Mismatch Negativity, Speaker Identity

## I. INTRODUCTION

The multimodality of everyday language comprehension and production is immediately obvious. Even as infants, our experience with language typically involves the processing of both auditory and visual information (Kuhl and Meltzoff, 1982; 1984). Indeed, research has established that listeners can use visual cues from the speaker's face to improve discrimination of speech sounds, e.g., discrimination of the phonemic contrast /p/ – /k/ is facilitated by simultaneous presentation of the visually discriminable mouth movements required to make these sounds (Sumby and Pollack, 1954; Liberman, et al., 1967; McGurk and MacDonald, 1976; Kuhl and Meltzoff, 1982; 1984; Hollich et al., 2005, Hazan et al., 2005, 2006). The speaker's face, however, provides more information than just the mouth movements used in producing speech. In particular, it provides an indexical cue of the identity of the speaker. The current paper examines whether there is separable encoding of the identity of the speaker in auditory speech perception and whether speech perception, in particular, phoneme discrimination, is influenced by visual speaker identity: Does consistent pairing of different faces with different sounds, i.e. hearing one speaker say one sound, and a second speaker saying a different sound, influence later auditory-only adult discrimination of the sounds?

There are a number of reasons to argue for an influence of speaker identity on speech perception. First, audio-visual models of speech processing argue for separable encoding of speaker identity in speech processing (Belin et al., 2004; Von Kriegstein et al., 2008). According to these models, speaker identity (visual or auditory) forms a necessary link between speech recognition and speaker recognition with information about the speaker's voice and the speaker's face being freely shared between auditory voice areas and visual face areas (Von Kriegstein et al., 2008). Similarly, exemplar models argue for a direct influence of (auditory) speaker identity

on auditory recognition (e.g., Johnson, 1990; Goldinger, 1990; Pisoni, 1990). The underlying assumption of such models is that each token of auditory input is stored along with information of speaker identity. Hearing the auditory token triggers all the stored information and makes it easier to discriminate the heard token from another token. An extension to visual speaker identity could easily be incorporated into such models. Having different visual speaker identity information attached to different auditory tokens should, similarly, make it easier to discriminate between the two auditory tokens.

Indeed, Von Kriegstein and colleagues document evidence of a special relationship between the speech produced by a speaker and the speaker's face (von Kriegstein et al., 2005; Calvert et al., 1997). Thus hearing a person's voice activates the fusiform faces area (typically associated with face processing) while videos of an articulating face activate the auditory cortex. Furthermore, seeing a person's face improves recognition of this person's voice (Schweinberger et al., 1997).

Second, auditory speaker identity (established through the speaker's voice) has a robust influence on phonemic and lexical processing. Recent research reports that recognition of words decreases when the voice of the speaker changes (Creelman, 1957; Mullenix et al., 1989); being able to identify the talker improves recognition of novel words (Nygaard et al., 1994); discrimination of vowels differs based on changes to perceived speaker identity (Johnson, 1990); and that perception of a sound ambiguous between /s/ and /f/ is influenced by higher-level knowledge of the identity of a speaker (Kraljic and Samuel, 2005; Eisner and McQueen, 2005). These studies suggest the real possibility of independent encoding of speaker identity (established through the speaker's voice) and auditory-linguistic content (the phonemic or lexical content of the message) as well as an influence of auditory speaker identity on speech perception.

The current paper extends this to ask whether there is similar influence of visual speaker identity (the speaker's face) on the brain's discrimination of phonetic categories.

Indeed, recent work by Mitchel and Weiss (2010) finds that adults are better able to discriminate two artificial language speech streams if the languages are consistently associated with two different visually presented faces. Here, adults were familiarised with two streams of artificial language such that one stream is always presented by one speaker and the other stream is presented by a distinct speaker. Adults were better able to recognise words from these artificial language speech streams when presented with consistent familiarisation (i.e., consistent language-speaker pairings relative to inconsistent language-speaker pairings). Von Kriegstein et al. (2008) term such improvement in auditory processing due to the concurrent visual presentation of the speaker's face "face-benefit", i.e., an improvement in auditory speech recognition due to previous exposure to linguistically non-relevant facial information, in particular when the target auditory stimuli to be discriminated are previously distinctly associated with discriminable faces.

The current study extends this work in two ways. First, we examine whether the reported face-benefit in adult speaker recognition or speech-stream discrimination can be found also in lower levels of processing, i.e., phoneme discrimination. That is, we ask whether associations between speaker identity and speech exist only at the higher levels of speech processing (discriminating between two languages) or whether they impact even bottom-up processing (phoneme discrimination). Second, we examine the time-course of an effect of visual speaker identity on speech processing using an auditory oddball event-related potential (ERP) task and not offline tests such as word-identification employed in previous studies (Mitchell and Weiss, 2010). The behavioural response is the end-product of a number of processes, e.g., word

recognition, button-press, in contrast to implicit monitoring of the brain's immediate response to an auditory stimulus. It is, therefore, easier to isolate the stage at which a stimulus change impacts processing (with millisecond accuracy) using ERPs relative to behavioural tasks. This is particularly well suited to the study of rapidly changing stimuli such as speech stimuli.

### **The current study**

The current study presented German adults with a series of familiarisation videos where adults saw two speakers' faces consistently paired with two non-native Hindi speech sounds, e.g., they saw Speaker 1 saying the Hindi dental sound /ḍa/ and Speaker 2 saying the Hindi retroflex sound /ɖa/. This contrast is particularly difficult for non-native speakers of Hindi to differentiate (Werker and Tees, 1984) – there should, therefore, be no difference in the brain activity to these two sounds when testing non Hindi-speaking populations, e.g., the German adults tested in the current study. The experiment was divided into a pre-familiarisation auditory-only oddball task, a familiarisation phase, and a post-familiarisation oddball test. In the oddball tests presented to participants in the pre- and post-familiarisation phase, adults were tested on their detection of the change from one sound to the other. In the familiarisation phase sandwiched by tests of phoneme discrimination, adults were presented with the familiarisation videos described above. We examined whether the brain's detection of the change from one sound to the other improves from the pre- to the post-familiarisation test, thereby showing a reliable influence of the familiarisation phase on speech-sound discrimination.

It is possible, however, that the introduction of the faces improves change detection regardless of the consistency of the pairing, perhaps by making the stimuli more interesting or by giving adults greater exposure to the auditory stimuli. In order to establish an influence of visual

speaker identity on speech discrimination, it is necessary to show that an improvement in change detection is driven by amodal information pertaining to the consistency with which the speakers' faces were paired with the sounds. Therefore, a control group of adults were presented with an inconsistent familiarisation phase, where they saw both speakers producing both /ɖa/ and ɖa/. Consequently, across participants familiarised with consistent and inconsistent face-sound pairings, exposure to auditory and visual stimuli as well as pre- and post-familiarisation phoneme discrimination tests was identical – the only difference was the consistency of pairing of the auditory and visual stimuli in familiarisation.

Phoneme discrimination in the pre- and post-familiarisation phase was tested using an auditory-only ERP oddball task, where participants were presented with Hindi /ɖa/ and ɖa/ tokens. As is standard in the auditory oddball task, one of the tokens, the standard, was presented with a higher frequency compared to the other, the deviant, with 80% - 20% split between standards and deviants. Change detection in oddball tasks is characterised by a mismatch-negativity between 150ms to 250ms in frontal electrodes, with more negative ERPs to rarely presented deviants compared to frequently presented standards (Näätänen et al., 1978). This response is associated with the involuntary triggering of attention to the change (i.e., presentation of the deviant) from the regular acoustic background created by repeated presentation of the standard token. A number of studies have used the auditory oddball paradigm to investigate adult discrimination of speech sound categories and highlight the reliability of the associated mismatch negativity response with regard to speech sound discrimination (see Cheour, 2007 for a review of previous studies). Analysis will focus on fronto-central electrode sites in the time-window specific to reported mismatch responses, i.e., between 150 to 250ms post-stimulus onset. If consistent pairing of faces with sounds drives discrimination of these sounds even in the absence

of visual stimuli, then we should expect to find increased negativity to the deviant rarely-presented token only in the post-familiarisation test in participants exposed to consistent face-sound pairing. Given that we test discrimination of a non-native contrast, participants should find it difficult to detect the change from standards to deviants in all other phases i.e., pre-familiarisation test phase for participants exposed to consistent and inconsistent familiarisation and post-familiarisation test phase for participants exposed to inconsistent familiarisation.

## **II. EXPERIMENT**

### **Participants**

Fifty-two German adults aged between 19 and 28 years (mean age: 20.98 years) took part in the experiment after giving written informed consent. Two participants had to be excluded from the analysis for providing fewer than 10 trials per condition following artifact rejection (7 and 9 trials each). Participants had no exposure to Hindi and had normal hearing and normal/corrected vision. Participants were given course credits in the Psychology program as reimbursement for their time.

### **Materials**

#### *Auditory stimuli*

The auditory stimuli used in the current experiment were the Hindi /ɖa/ and /ɗa/ tokens. Auditory stimuli were recorded by a female native speaker of Hindi. Thus the same speaker produced the tokens paired with both faces. This ensured that we manipulated the consistency with which speakers' faces correlated with auditory tokens, but not speakers' voices. Five tokens of each sound were chosen to be paired with the visual stimuli for the audio-visual



familiarisation phase. In addition, one token of each sound was chosen for the pre- and post-familiarisation auditory test phase. To ensure that the stimuli were not easily discriminable (due to differences in the vowels of the two tokens), the vowel from the /ḍa/ token was spliced onto the vowel of the /da/ token. The stimuli were matched in intensity, but splicing the vowel from one token to the other led to differences in the duration of the stimuli due to naturally occurring differences in the voicing lag of the two tokens. The durations of the /ḍa/ and /da/ tokens for test were 541 and 428ms respectively. Note that whilst these differences might aid differentiation of the tokens, this should be identical across participants familiarised with consistent and inconsistent face-sound pairings and across pre- and post-familiarisation tests.

### ***Visual stimuli***

The visual stimuli were videos of two Caucasian females producing the German /da/ sound (similar to the Hindi /ḍa/) against a grey background. As in Mitchel and Weiss (2010), the speakers were specifically chosen for the differences in their appearance – see Figure 2 for a still from the videos of each speaker as well as a schematic of the experiment.

### ***Pairing of audio and visual stimuli***

Ten video-clips of each speaker were paired with five /ḍa/ and five /da/ auditory tokens, such that the auditory and visual stimuli were synchronised. This gave us five audio-visual /ḍa/ video-clips for Speaker 1, five audio-visual /ḍa/ video-clips for Speaker 2, five audio-visual /da/ video-clips for Speaker 1 and five audio-visual /da/ video-clips for Speaker 2. The twenty video-

clips were then grouped together to form two consistent and two inconsistent familiarisation videos. In the consistent pairing of faces and sounds, video-clips presented each speaker consistently saying a particular sound. Therefore, one consistent familiarisation video sequentially presented video-clips of Speaker 1 saying /dɑ/ and Speaker 2 saying /dɑ/ and the other consistent familiarisation video presented video-clips of Speaker 1 saying /dɑ/ and Speaker 2 saying /dɑ/. In the inconsistent videos, video-clips presented each speaker saying both sounds. For instance, any inconsistent familiarisation video presented video clips of Speaker 1 saying /dɑ/, Speaker 2 saying /dɑ/, Speaker 1 saying /dɑ/, Speaker 2 saying /dɑ/. Overall, we created two consistent videos and two inconsistent videos with 20 face-sound pairings per video. We ensured that within the 20 face-sound pairings, participants received equal exposure to /dɑ/ and /dɑ/ tokens as well as to Speaker 1 and Speaker 2. Each familiarisation video was exactly 58 seconds long with 500ms of a black screen between each video-clip. Note that the physical video-clips presented to participants across consistent and inconsistent familiarisation videos were identical – the difference was that consistent videos consisted of consistent speaker-sound pairings within subjects and inconsistent videos consisted of all possible pairings of speakers and sounds.

## **Procedure**

Half the adults received consistent familiarisation and the other half received inconsistent familiarisation. The experiment was divided into a pre-familiarisation phoneme discrimination test, a familiarisation phase, and a post-familiarisation phoneme discrimination test. Prior to and post familiarisation, an auditory-only odd-ball task tested adults' automatic detection of the

change from one sound to the other – participants were not presented with the corresponding video tokens during test.

### ***1. Pre- and post-familiarisation phoneme discrimination tests***

Across participants, these were identical to one another. Here, participants were presented with 500 auditory tokens with 400 repetitions of the standard token and 100 repetitions of the deviant token (80% standard, 20% deviant). For half of the participants (split across both kinds of familiarisation groups), /d̥a/ was the standard and d̥a/ was the deviant. For the other half, /d̥a/ was the standard and /d̥a/ was the deviant. Order of presentation was pseudo-randomised with 4 deviant tokens distributed across every 20 trials. The inter-stimulus-interval between any two tokens varied between 550 to 700ms at 550, 600, 650 or 700ms. The pseudo-random variation in ISI might make the tokens more difficult to discriminate, camouflaging the voicing lag between the /d̥a/ and /d̥a/ tokens. During pre- and post-familiarisation tests, participants were presented with repetitions of the auditory stimuli as they watched a silent film of their choice.

### ***2. Familiarisation phase***

Once participants completed the pre-familiarisation discrimination test, their attention was directed to the screen in front of them where familiarisation videos were presented. Each participant received a total of six repetitions of either a consistent or an inconsistent familiarisation video on a loop, thereby receiving a total of 464 seconds (7.73 minutes) of familiarisation.

Once participants completed familiarisation, they were presented with the continuation of their movie and the post-familiarisation phoneme discrimination test. Across participants exposed to consistent familiarisation, we counter-balanced the pairing of speakers with sounds during familiarisation as well as the first speaker-sound pairing presented to participants. Participants exposed to inconsistent familiarisation were presented with all possible pairings of speakers and sounds. The pre- and post-familiarisation phoneme discrimination test were not only identical to each other but were also identical across all participants. Therefore, any differences in the ERPs to pre- and post-familiarisation across participants exposed to consistent and inconsistent face-sound pairings can only be attributed to the familiarisation phase and to the consistency of face-sound pairings, in particular (since the video-clips used to form consistent and inconsistent videos were identical across participants).

### ***3. Electrophysiological recording and data analysis***

Electrophysiological data was recorded using the Biosemi Active Two Amplifier system at a sampling rate of 2048 Hz from 32 Ag/AgCl electrodes placed according to the 10-20 convention. Electrode offsets were kept  $< 25 \mu\text{V}$ . Offline analysis of the continuous EEG data was conducted using the Brain Electrical Source Analysis package (BESA). Electroencephalogram was re-referenced offline to the averaged mastoid reference. Epochs were defined from -100 to 800 ms from the onset of the auditory token. EEG data was filtered off-line using a 0.1 Hz high-pass forward filter and a 20 Hz low-pass, zero-phase shift filter. Blink and movement artifacts were automatically rejected using a 120 Hz amplitude cut-off across eye electrodes across the entire epoch. Baseline correction was performed in reference to pre-stimulus activity (-100 to 0 ms). Given the known onset and scalp distribution of the mismatch

negativity response (Näätänen et al., 1978), analysis will focus on fronto-central electrode sites in the time window between 150ms to 250ms. For analysis, we analysed data from frontal left (AF3, F3), frontal right (AF4, F4) and frontal and central midline electrode sites (Fz, Cz) to compute mean activity across fronto-central sites separately for standards and deviants across participants exposed to consistent and inconsistent pairings in the pre- and post-familiarisation phoneme discrimination test phase. Preliminary analysis revealed no significant interaction between electrode site or laterality with condition ( $ps > .2$ ), so subsequent analyses collapsed the data across all fronto-central electrode sites.

### III. RESULTS

Figures 1 and 2 plot the event-related potentials to standards and deviants in the pre- and post-familiarisation phase for subjects exposed to consistent (Figure 1) and inconsistent familiarisation (Figure 2) averaged across all frontal electrode sites. A mixed factor ANOVA with *phase* (pre-, post-familiarisation) and *condition* (standard, deviant) as within-subjects factors and *familiarisation type* (consistent, inconsistent familiarisation) as a between-subjects factor found a significant interaction between phase, condition and familiarisation type,  $F(1, 48) = 5.06$ ,  $p = .029$ ,  $\eta_p^2 = .095$ . ANOVAs with phase and condition as within-subjects factors separately performed for participants exposed to consistent and inconsistent familiarisation found a significant interaction between phase and condition for those participants exposed to consistent familiarisation ( $F(1, 24) = 4.09$ ,  $p = .058$ ,  $\eta_p^2 = .14$ ) but not for those participants exposed to inconsistent familiarisation ( $F(1, 24) = 1.4$ ,  $p = .24$ ).

There was a significant difference in the brain activity to standards and deviants across all fronto-central electrode sites in the post-familiarisation phase,  $F(1, 24) = 7.16$ ,  $p = .013$ ,  $\eta_p^2 =$

.23, but not in the pre-familiarisation phase,  $F(1, 24) = 1.24$ ;  $p = .2$ , in participants exposed to consistent familiarisation. In contrast, there was no significant difference in the brain activity to standards and deviants in the pre-familiarisation phase,  $F(1,24) = 1.06$ ,  $p = .3$ , or post-familiarisation phase,  $F(1, 24) = .25$ ,  $p = .6$ , in participants exposed to inconsistent familiarisation. As expected, German adults in both familiarisation groups did not detect the change from one phoneme to the other prior to familiarisation. However, analysis of the post-familiarisation phoneme discrimination test confirmed the brain's detection of the phoneme change and discrimination of the non-native sounds only in those participants exposed to consistent face-sound pairings during familiarisation.

-----  
INSERT FIGURES 1 AND 2 ABOUT HERE  
-----

Exposure to consistent pairing of faces and sounds during familiarisation improved the brain's automatic discrimination of a difficult non-native speech sound contrast, thereby arguing for a strong and early influence of visual speaker identity (established through the speakers' face) on speech sound discrimination. The current study presents the first evidence that amodal information pertaining to the consistent pairing of speakers (i.e., their faces) and speech sounds can support adults' learning of new phonetic categories (see Figure 3).

-----  
INSERT FIGURE 3 ABOUT HERE  
-----

#### **IV. GENERAL DISCUSSION**

A vigorous debate in the field of speech perception concerns the integration of visual cues in auditory speech processing. On the one hand, auditory-only models of speech processing argue for the relative independence of auditory processing from the processing of visual cues (Hickok and Poeppel, 2007) such that auditory processing involves the processing of only auditory information. On the other hand, audio-visual models of speech perception argue for early integration of visual cues in auditory processing (e.g., Belin et al., 2004; Von Kriegstein et al., 2008), such that visual cues can influence even the early stages of auditory processing. Indeed, some audio-visual models (Belin et al., 2004; Von Kriegstein et al., 2008) suggest that speaker identity (visual or auditory) forms a necessary link between speech recognition and speaker recognition. The results of the current study strongly support an influence of visual speaker identity on speech perception.

The notion of acquired distinctiveness (Miller and Dollard, 1941; Hall, 1991) explains such effects neatly: the concurrent presentation of auditory and visual information provides an additional source of information for listeners to exploit in speech discrimination. In the current study, for instance, there is a lot of information discriminating not just the two sounds but also different tokens of the two sounds. The addition of a salient and consistent visual cue (the faces of the speakers) distinctly paired with the sounds (Face 1 with Sound 1 and Face 2 with Sound 2) can help listeners focus on the differences relevant to discriminating between the sounds and ignore differences between different tokens of the same sound.

Can the results of the current study conclude in favour of a special role for visual speaker identity in speech processing? Mitchel and Weiss (2010) argue in favour of such a special role with their finding that adults are better able to discriminate two artificial language speech

streams only if the languages are consistently associated with two different visually presented faces but not if the two languages are consistently associated with two distinctly coloured backgrounds. At present, the current study can only conclude that visual speaker identity is one of potentially numerous visual cues that influence auditory speech processing (see e.g., Yeung and Werker, 2009; Cunillera et al., 2010; Hayes-Harb, 2007; Teinonen, Aslin, Alku, & Csibra, 2008). That is, our study suggests that one visual cue available to listeners in communication is the identity of the speaker, and that listeners readily use such cues to improve their categorisation of speech sounds. This might be used by adults in tapping into the acoustic-phonetic space characteristically used by a particular speaker – seeing a speaker’s face may then help pre-activate phonetic cues specific to this speaker and aid speech processing. Given Mitchel and Weiss’s finding of a benefit in speech processing attributable solely to visual speaker identity, it would be interesting to examine whether there is a similarly special role for visual speaker identity in phoneme discrimination.

Relatedly, one could ask whether a distinct association between speaker identity and speech perception is required for efficient speech processing. Previous studies would suggest this not to be the case, e.g., studies showing that misaligned speaker gender information (female voices, male faces) does not impact the magnitude of the McGurk effect (Green, Kuhl, Meltzoff, & Stevens, 1991). Taken together with the current results, this would suggest a non-essential but, nevertheless, influential role of visual speaker identity in phonetic category learning that might be used by listeners to keep track of speaker-specific phonetic cues in speech processing.

For instance, this influence of visual speaker identity on speech processing might constitute an important cue for bilinguals to separate their two languages in the mind. Consistently associating distinct sounds with distinct speakers might help bilinguals to trigger the appropriate



language immediately upon seeing speakers of that language. Indeed, bilinguals are exposed not just to different people speaking different languages, but equally to different people speaking different sounds. Even when two languages share the same sound, e.g., the sound /d/ which occurs in both English and French, there are systematic differences in the acoustic characteristics of this sound when produced in English and French by English-French bilinguals (Sundara, 2005). Furthermore, bilingual adults are sensitive to such subtle phonemic cues in spoken language processing (Ju and Luce, 2004). By examining the basic levels of language processing, i.e., phoneme discrimination, the current study suggests one mechanism bilinguals might use to separate their two languages. Associating sounds with people might help bilinguals to *prioritise* words from one of their languages and speed language processing.

This finding also has implications for our understanding of the interaction between voice recognition and face recognition in speech processing. A number of studies suggest that we are faster to recognise a speaker's voice when primed by the speaker's face (e.g., Schweinberger et al., 1997, 2007). The results of the current study establish one mechanism for such cross-modal effects of speaker identity. That is, by associating a particular speaker with certain sounds or certain pronunciations of sounds (indeed, to a German native speaker, the two sounds presented should normally sound like variant tokens of the German /d/), we might begin to map the characteristics of a speaker's voice with their face. It would be interesting to see whether there are further facilitatory effects of voice-face correlations in language processing – for e.g., are we faster to recognise words spoken by a speaker when primed by this speaker's face as opposed to an unrelated speaker's face? Alternatively, one could examine the interaction between speaker voice and visual speaker identity in driving phonetic learning given that the current study explicitly controlled for effects of speaker voice.

Furthermore, developmentally, there is a simultaneous narrowing of our perceptual sensitivities to own face and own language stimuli: By 9 months of age, infant face processing has achieved a state of maturity, with infants showing good discrimination of own-race faces and being unable to discriminate other-race faces (Pascalis et al., 2002; Kelly et al., 2007). Notably, this selective attention to own-race differences in 9-month-olds closely parallels the age at which infants show a selective attention to own-language speech sounds compared to non-native speech sounds (Werker and Tees, 1984). This simultaneous narrowing of speech-specific and face-specific sensitivities suggests a potential link between the two processing abilities that is in complete agreement with the results of the current study (Weikum et al., 2007; Pons et al., 2009). It would, therefore, be interesting to examine whether similar effects of visual speaker identity exist in infancy.

Finally, we have shown here that visual speaker identity influences speech processing early on, i.e., 150ms after the onset of a sound. As noted above, most previous studies have examined the influence of speaker identity on language processing (not voice recognition) using behavioural tasks such as word-identification employed in previous studies (Mitchell and Weiss, 2010). The finding of early discrimination of the two sounds, in the absence of any visual cues, might be taken to suggest that exposure to consistent speaker-sound pairings fine-tunes our representation of sounds by drawing two sound categories further apart in the listener's perceptual space – the MMN is typically associated with the involuntary (pre-attentive) triggering of sensitivity to the difference between two sounds. Such warping of auditory-perceptual space is typically noted in research with infants, showing that auditory exposure distorts the perceptual distance between two sound categories such that two acoustically similar sounds are perceived as categorically distinct from one another (Kuhl et al., 2006; Kuhl, 2004).

The current research extends this to suggest that audio-visual exposure, in particular, linguistically non-relevant audio-visual exposure, has a similar effect of distorting a listener's auditory perceptual space.

## **V. CONCLUSIONS**

The current study presents strong evidence of the influence of visual speaker identity on speech processing. Whilst previous studies have shown effects of speaker identity on language discrimination or speaker recognition, the current study extends this to demonstrate a robust effect of visual speaker identity on the building blocks of speech processing, i.e., phonetic category learning. The finding that such cross-modal effects apply even in basic speech-sound discrimination suggest modality-nonspecific retrieval of information from the environment guiding bottom-up speech processing – as speech comprehenders, we appear to use all the information that is available to us, visual and auditory alike, linguistic and non-linguistic alike, to optimise our processing of speech.

## REFERENCES

- Belin, P., Fecteau, S., and Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception, *Trends in Cognitive Science*, 8, 129-135.
- Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R. and McGuire, P. K. (1997). Activation of auditory cortex during silent lipreading, *Science*, 276, 593-596.
- Creel, S. C., Aslin, R. N., and Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access, *Cognition*, 106, 633–664.
- Cunillera, T., Càmarà, E., Laine, M., and Rodríguez-Fornells, A. (2010). Speech segmentation is facilitated by visual cues, *Quarterly Journal of Experimental Psychology*, 63, 260-274.
- Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *Neuroreport*, 8(4), 919–924.
- Eisner, F., and McQueen, J. M. (2005). The specificity of perceptual learning in speech processing, *Perception and Psychophysics*, 67(2), 224-238.
- Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Perception & psychophysics*, 50(6), 524–536.
- Goldinger, S. (1990). Effects of talker variability on self-paced serial recall, *Research on Speech Perception*. Bloomington: Indiana University Press.
- Hall, G. F. (1991). *Perceptual and Associative Learning*. Oxford, UK: Clarendon Press.
- Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, 23(1), 1-31

- Hazan, V., Sennema, A., Iba, M., and Faulkner, A.,(2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, 47, 360-378.
- Hazan ., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., and Chung, H. (2006). The use of visual cues in the perception of nonnative consonant contrasts, *Journal of the Acoustical Society of America*, 119, 1740-1751.
- Hickok, G., and Poeppel, D. (2007). The cortical organisation of speech processing, *Nature Reviews Neuroscience*, 8, 393-402.
- Hollich, G., Newman, R. S., and Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech, *Child Development*, 76, 598-613.
- Ju, M., and Luce, P.A. (2004). Falling on sensitive ears: Constraints on Bilingual Lexical Activation, *Psychological Science*, 15, 315-318.
- Johnson, K. (1990). Contrast and normalization in vowel perception, *Journal of Phonetics*, 18, 229-54.
- Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge., L., and Pascalis, O. (2007). The other-race effect develops during infancy: evidence of perceptual narrowing, *Psychological Science*, 18, 1084-1089.
- Kraljic, T. and Samuel, A.G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51, 141-178.
- Kuhl, P.K. and Meltzoff, A.N. (1982). The bimodal perception of speech in infancy, *Science*, 218, 1138-1141.
- Kuhl, P.K. and Meltzoff, A.N. (1984). Intermodal speech perception, *Infant Behavior and Development*, 7, 361-381.

- Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code, *Nature Reviews Neuroscience*, 5, 831-843.
- Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months, *Developmental Science*, 9, F13-F21.
- Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967). Perception of the speech code, *Psychological Review*, 74, 431-61.
- McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices, *Nature*; 264, 746-748.
- Miller, N. E., and Dollard, J. (1941). *Social learning and imitation*. New Haven, CT: Yale University Press.
- Mitchel, A., and Weiss, D.J. (2010). What's in a face? Visual contributions to speech segmentation, *Language and Cognitive Processes*, 25, 456-482.
- Mullennix, J.W., and Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception, *Perception and Psychophysics*, 47, 379-390.
- Näätänen, R., Gaillard, A. W. K., and Mäntysalo, S. (1978). Early selective attention effect on evoked potential reinterpreted, *Acta Psychologica*, 42, 313-329.
- Nygaard, L.C., Sommers, M., and Pisoni, D.B. (1994). Speech perception as a talker contingent process, *Psychological Science*, 5, 42-46.
- Pascalis, O., de Haan, M., and Nelson, C.A. (2002). Is face processing species specific during the first year of life, *Science*, 296, 1321-1323.
- Pisoni, D. (1990). Effects of talker variability on speech perception: implications for current research and theory, *Proceedings of the 1990 International Conference on Spoken Language Processing*, 1399, 407. Tokyo: The Acoustical Society of Japan.

- Pons, F., Lewkowicz D. J., Soto-Faraco S., & Sebastian-Galles N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of America*, 106, 10598-10602.
- Schweinberger, S. R., Herholz, A. and Stief, V. (1997). Auditory long-term memory: Repetition priming of voice recognition, *Quarterly Journal of Experimental Psychology*, 50A, 498-517.
- Sumbly, W.H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise, *Journal of Acoustical Society of America*, 26, 212-215.
- Sundara, M. (2005). Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French, *Journal of the Acoustical Society of America*, 118, 1026-1037.
- Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, 108(3), 850–855.
- von Kriegstein, K., Kleinschmidt, A., Sterzer, P. and Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition, *Journal of Cognitive Neuroscience*, 17, 367-376.
- von Kriegstein, K., Dogan, O., Grüter, M., Giraud, A. L., Kell, C., Grüter, T., Kleinschmidt, A., and Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition, *Proceedings of the National Academy Sciences*, 105, 6747-6752.
- Weikum, W. M., Vouloumanos A., Navarra J., Soto-Faraco S., Sebastián-Gallés N., & Werker J. F. (2007). Visual language discrimination in infancy. *Science*, 316, 1159.

Werker, J.F. and Tees, R.C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life, *Infant Behavior and Development*, 7, 49-63.

Yeung, H.H. and Werker, J.F. (2009). Learning words' sounds before learning how words sound: 9-month-old infants use distinct objects as cues to categorize speech information, *Cognition*, 113, 234-243.



**FIGURES**

Figure 1: Event-related potentials (ERPs) to standards and deviants in the pre- and post-familiarisation phase for subjects exposed to consistent familiarisation. Graphs present data averaged across fronto-central electrode sites (AF3, AF4, F3, F4, Fz, Cz) plotted from -100 to 400ms from the onset of the sound (MMN window – 150 to 250ms – shaded in grey).

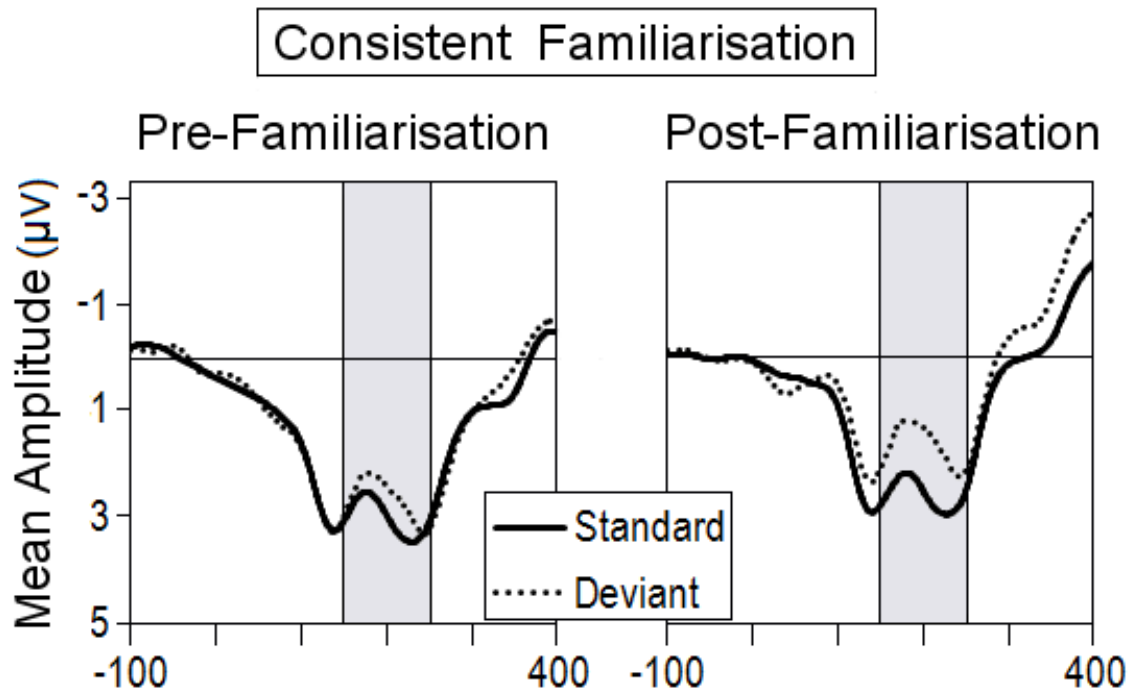


Figure 2: Event-related potentials (ERPs) to standards and deviants in the pre- and post-familiarisation phase for subjects exposed to inconsistent familiarisation. Graphs present data averaged across fronto-central electrode sites (AF3, AF4, F3, F4, Fz, Cz) plotted from -100 to 400ms from the onset of the sound (MMN window – 150 to 250ms – shaded in grey).

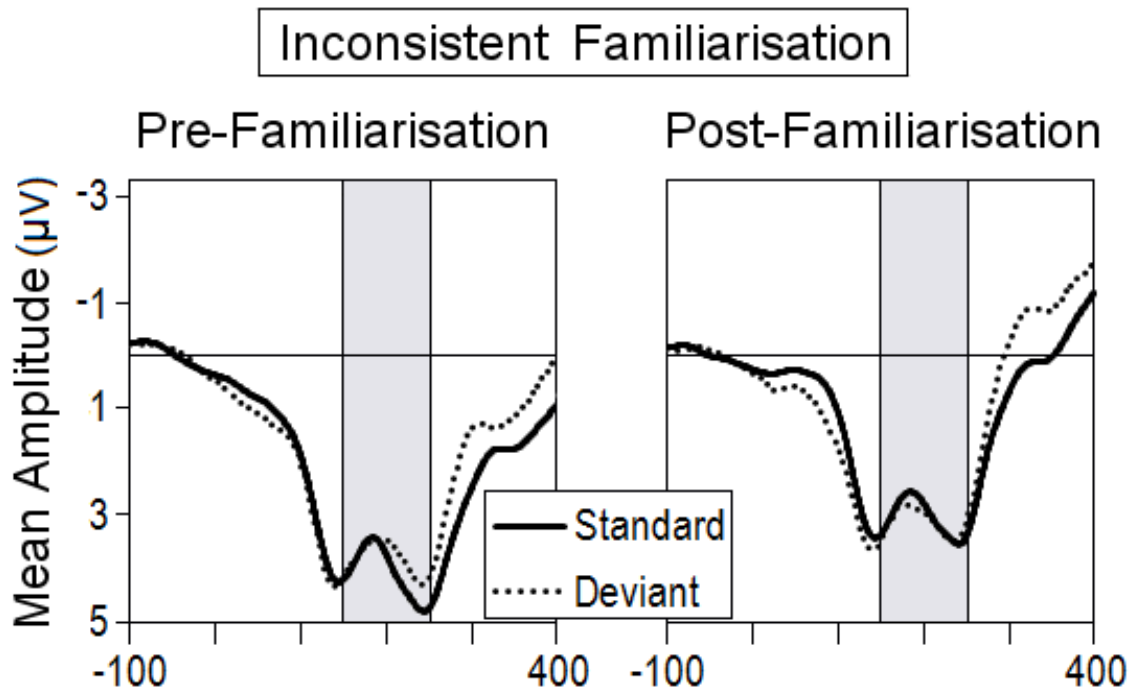










Figure 3: Schematic of Experiment and results

	Pre-familiarisation test	Familiarisation 6 Blocks of 20 random tokens				Post-familiarisation test	
<b>Consistent speaker-sound familiarisation</b>	<u>Auditory oddball task</u> 400 repetitions of Sound 1 and 100 repetitions of Sound 2	Visual stimuli	 Speaker 1	 Speaker 2	 Speaker 1	 Speaker 2	<u>Auditory oddball task</u> 400 repetitions of Sound 1 and 100 repetitions of Sound 2
	No discrimination of sounds	Auditory stimuli	da Sound 1	ɔa Sound 2	da Sound 1	ɔa Sound 2	<b>Successful discrimination of sounds</b>
<b>Inconsistent speaker-sound familiarisation</b>	<u>Auditory oddball task</u> 400 repetitions of Sound 1 and 100 repetitions of Sound 2	Visual stimuli	 Speaker 1	 Speaker 2	 Speaker 1	 Speaker 2	<u>Auditory oddball task</u> 400 repetitions of Sound 1 and 100 repetitions of Sound 2
	No discrimination of sounds	Auditory stimuli	da Sound 1	ɔa Sound 2	ɔa Sound 2	da Sound 1	No discrimination of sounds