INFLUENCES OF VISUAL SPEECH INFORMATION ON THE PERCEPTION OF FOREIGN-ACCENTED SPEECH IN NOISE

Saya Kawase, Jeesun Kim, Vincent Aubanel, Chris Davis

The MARCS Institute, University of Western Sydney
s.kawase@uws.edu.au

ABSTRACT

This study examined the extent to which visual speech assisted native Australian English speakers to perceive Japanese-accented English compared with Australian English spoken sentences presented in speech-shaped noise (SNR:-4dB). Twenty-one native Australian English listeners performed a speech perception in noise task with Japanese-accented and Australian English sentences in two conditions: an Audio-only condition (AO) and an Audio-visual condition (AV) where the talker's face was also shown. The results showed that the addition of visual speech information facilitated the speech perception for both Australian English and Japanese-accented English. However, the visual benefit was significantly smaller in the perception of Japanese-accented English, indicating that foreign-accented speech affects the visual speech benefit potentially due to non-native visual form and timing differences.

Keywords: foreign-accented speech, audio-visual speech perception.

1. INTRODUCTION

It is well established that seeing a talker's face/head movements (visual speech) helps speech perception, particularly in degraded listening environments such as in the presence of background noise [12, 14, 16]. Similar to auditory speech, the information transmitted by visual speech can be described at the segmental and non-segmental levels (roughly *form* and *timing* information [10]). Visual form information relates to changes in articulatory gestures (with mouth, lip, tongue) and can provide information about speech segments as phonemes (although this is more coarse-grained than that provided by auditory information).

The effect of form information from visual speech has been extensively examined with respect to phonemic perception. For instance, the McGurk effect is created by presenting both auditory and visual information where the perceptual integration of visual /g/ and auditory /b/ cues result in the perception of /d/ [11]. Visual timing cues (e.g., the cyclic opening and closing of the mouth and perioral regions) also provide useful information about speech onset, intensity and offset and in doing so provide rhythmic information [3, 4, 7]. Indeed, it has been reported that young infants can discriminate different rhythm languages (e.g., English vs. French) based on visual speech alone, suggesting the availability of language rhythm information in speech movements [18]. It has also been suggested that visual timing information can help segment speech so that speech can be more accurately recognized; particularly in noise [4].

Given that the vast majority of the previous studies have used native speech stimuli (tokens or continuous speech produced by a native speaker), the extent to which foreign-accented visual speech will facilitate speech perception is unclear. For instance, the visual form and timing information produced by non-native speakers may be different to that produced by native ones because of the influences of the non-native speaker's native language. While extensive auditory-based research has demonstrated a native language influence on non-native speech production (e.g., the speech learning model (SLM); Flege [5]), very few studies have investigated whether there is a similar influence on visual speech production. Such an influence from a native language is expected since, for example, language-specific articulatory settings exist and language rhythms vary across languages [18, 19]. What is not clear, however, is whether native and non-native articulation will be visibly different, and if so, whether such a difference will affect speech perception.

A recent study illustrates that non-native visual speech form can have a negative influence on auditory speech processing [9]. In this study, native English listeners were asked to identify English CV syllables produced by Japanese learners of English and native English speakers in audio-visual (AV), audio-only (AO) and visual-only (VO) conditions. The results showed that seeing non-native Japanese speakers of English produce visually salient consonants (/b, v, θ/) was facilitatory. In contrast, an inhibitory effect was found for Japanese-produced /ɹ/: whereby seeing the talker produce this token resulted in lower intelligibility compared to when the listener could not see the talker. A follow-up analysis suggested that this negative visual effect

may be due to the Japanese speakers' different articulation (i.e., lack of lip-rounding), leading to native English perceives to identify /la/ instead.

In contrast to [9], [8] did not find any non-native talker's visual speech effects. In [8], native English listeners were asked to decide if each sentence produced by non-native English speakers was true or false by pressing T or F on a keyboard. The task was conducted in AO and AV conditions, but there was no significant difference in their performance regardless of the availability of visual speech. The inconsistent results may be due to differences between the studies. First, [9] used syllables whereas [8] used sentences. For syllable production, any native language rhythmic influence is likely to be minimal; whereas such an influence may be much greater in sentence production. One possibility then, is that the null visual speech effects found in [8] was due to the interfering effects of non-native visual speech timing information. However, some caution needs to be exercised before adopting this interpretation. This is because the stimulus sentences in [8] were presented in quiet, so participants would be less likely to rely on visual speech (c.f., [17]) and so no visual speech effect was found. Given this, a further examination using sentence material is required; one in which non-native sentences are presented in noise for speech identification, as in [9].

The current examination was thus designed to investigate the influences of visual speech (both form and timing) on the perception of Japanese accented English in noise. In the experiment, native English participants were presented with two groups of talkers: Japanese learners of English and native English talkers, in AO (with a static figure; no mouth movement) and in AV (with mouth movement) conditions. It was expected that there would be visual facilitative effects as shown in previous studies [3, 4, 7]. However, it was also expected that there would be a reduced facilitative or no visual effect on the perception of speech in Japanese accented English compared to native English speech due to the influence of non-native visual form and timing [8, 9].

2. EXPERIMENT

2.1. Participants

Twenty one native Australian English perceivers (sixteen female, five male; M_{age}= 24.0 years) participated in this study. They were recruited from the University of Western Sydney using the university's research participation system. All of the participants reported normal hearing and normal or corrected-to-normal vision. Data from one female participant was excluded due to her language background (i.e., simultaneous bilingual).

2.2. Stimulus materials

The stimulus sentences (IEEE Harvard Sentences) consisted of 160 sentences. The stimuli were produced by two Japanese and two Australian English talkers (all female). The Japanese talkers were newly arrived Japanese learners of English (M_{age}= 29.0 years; M_{LOR}= 4.5 months). They started to learn English as a foreign language in Japan at approximately age 13. Their daily exposure to spoken English was limited and none of them used English at home or had lived in an English speaking country prior to their arrival in Australia. Thus, at the time of testing, they were considered late, intermediate level English learners (cf. [6]). The two Australian English talkers were postgraduate students at the University of Anonymity.

The audio and video recordings were made in a sound-treated recording booth. The Japanese and English talkers were given instructions regarding facial expression (neutral) and pose (forward facing, at camera). They were asked to read a list of sentences, one at a time, out loud in a neutral tone whilst being recorded. The set of sentences were recorded twice for each participant, but only the first production was used unless errors or disfluencies occurred in the first production. Each sentence was presented for participants to utter on a 17" LCD computer monitor using DMDX software. The videos were recorded using a Sony HXR-NX30P video camera.

Separate audio recording were made using an externally connected lapel microphone, (an AT4033a audio-technica microphone) in 44.1 kHz, 16-bit mono. These auditory signals were dubbed onto the video files. The auditory signals were mixed with the associated speaker's speech-shaped noise at a signal-to-noise (SNR) ratio of -4dB, using Praat software [2]. The speech shape noise was produced on the basis of the entire track of the talker's production and was added to the entire stimuli. For the video files, the location of the talkers' lips was tracked using Sensarea software [1] and the videos were edited to ensure that the stimulus talkers' face appeared in approximately the same position across trials. The edited audio and video files were synchronized with a tailored Matlab script based on Psychtoolbox (MATLAB R2013a).

Two types of stimuli were prepared: an AV condition (n=80) where the lower region of the face was presented with visible face and mouth motion, and an AO condition (n=80) where only a static face of the talker was presented. Each condition consisted

of the recordings from each of the four talkers (n=20 each), and none of the stimulus sentences were repeated.

2.3. Procedure

The participants were tested individually in a sound-treated booth. In the AV condition, they were instructed to watch the talker's articulatory movements while listening to the stimulus over the headphones; or in the AO condition to listen to the stimulus while viewing a static face. A set of MATLAB scripts based on Psychtoolbox were used for stimulus presentation and response collection. The two modality conditions (AV, AO) and two talker groups (Japanese, English) were counterbalanced across participants. In the task, the participants were asked to type in what they heard using a keyboard. In order to ensure that they watched the visual stimuli, additional catch trials (n=4) were presented in each condition. When a red cross appeared on a screen, they were asked not to respond to the sentence, but instead asked to press the enter button. Overall, the perception task lasted approximately 45 minutes.

3. RESULTS

The results of the catch-trials showed that one participant did not pay attention to the visual presentation. Data of this participant was removed and the data reported here is from the remaining 19 participants. The participants' mean correct identification data were analysed using a repeated measures ANOVA with talker group (Japanese and English) and stimulus presentation condition (AV, AO) as within-subject factors.

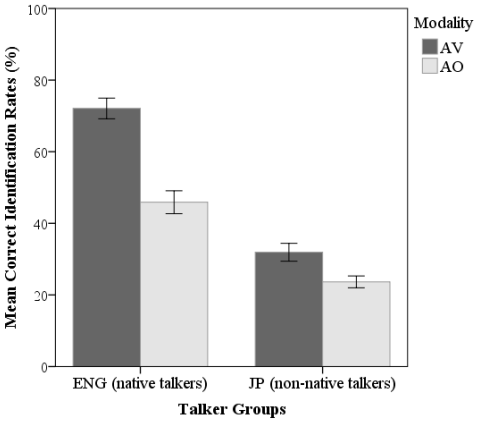

Figure 1: The mean correct identification rates (%) of English production produced by English (ENG) and Japanese (JP) talkers in the Audio-visual (AV) and the Audio-only (AO) conditions. Error bars indicate +/- one standard error.

Figure 1 shows the mean correct identification rates (%) of English sentences produced by native English (ENG) and Japanese (JP) talkers. Overall, the intelligibility rates of the Japanese-accented English (27.8%) was lower than the native English (59.0%) [$F(1, 18)=396.61$, $p<.001$, partial $\eta^2=.957$]. There was also a main effect for the stimulus presentation condition [$F(1, 18)=245.55$, $p<.001$, partial $\eta^2=.932$], showing that the overall intelligibility rates were higher in the AV (52.0%) compared to the AO (34.8%) condition. A significant interaction between the talker group and the stimulus condition was found [$F(1, 18)=61.36$, $p<.001$, partial $\eta^2=.773$]. Bonferroni adjusted *post hoc* tests revealed that the mean correct identification rates differed as a function of talker group and stimulus condition although both the intelligibility rates of both talkers, regardless of native and non-native talkers, were significantly higher in the AV compared to the AO ($p<.001$).

In order to show to what extent the visual speech information increased the intelligibility relative to the auditory intelligibility, the degree of visual enhancement (VE) was also measured using the formulae: VE= (AV-AO)/AO where AV and AO are the percentages of correct responses in audio-visual and audio-only conditions respectively.

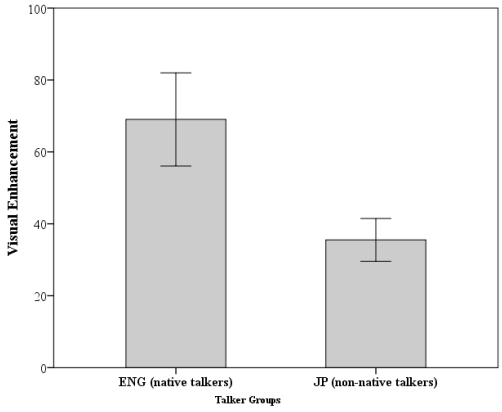

Figure 2: The mean visual enhancement of English production produced by English (ENG) and Japanese (JP) talkers. Error bars indicate +/- one standard error.

As can be seen in Figure 2, the visual enhancement was significantly lower for the Japanese–accented English (35.5) compared to the native English (69.0) [$t(18)=2.68$, $p<.05$]. Overall, the findings suggest that native English participants benefited more from the visual speech information when perceiving non-accented English compared to Japanese-accented English.

4. DISCUSSION

The aim of study was to investigate the extent that foreign-accented compared to native visual speech facilitates speech perception in noise. While visual speech benefit was widely observed in speech perception particularly in noise [12, 14, 16], little is known to what extent the visual speech benefits exist in foreign-accented speech perception. Given that non-native speech production is influenced by the native language (e.g., SLM [5]), it was expected to observe the less increase by the additional visual speech input compared to the auditory input only due to the non-native visual form and timing information.

The results showed that the degree of visual enhancement was significantly less for the Japanese-accented compared to native English speech. There are several possible explanations for this decreased visual enhancement effect with Japanese-accented speech. The most obvious explanation is simply that non-native visual form and timing cues were not as effective as the native ones. That is, the visual speech produced by Japanese talkers may be less able to reduce uncertainty in the auditory signal. This lack of precision may be due to lack of practice (in pronouncing English words) and/or because of the influence of Japanese articulatory production processes on the non-native English articulations. For example, in regard to this latter possibility, it has been suggested that talkers have language specific articulatory settings and that to produce native-like speech a talker has to use the language appropriate settings (e.g., [19]). Such settings could explain the non-native form effect that has been found previously in the perception of phonemes (e.g., [9]).

A related explanation is that in order to benefit from visual speech information a minimal level of AO perception is necessary. That is, native English perceivers may not have been able to use visual speech input efficiently due to the lower auditory intelligibility in the foreign-accented speech. It has been suggested that foreign-accented speech can be more easily degraded in noise compared to the native speech [13]. Thus although the amount of masking (i.e., noise level) was controlled, the degree of degradation in auditory intelligibility with noise may have been greater in the Japanese-accented English, potentially resulting in less effective audio-visual speech processing. Accordingly, the English perceivers may not be able to use the visual cues to the same extent as in their native speech perception.

These findings demonstrate the potential difficulties inherent in foreign-accented speech perception in face-to-face conversations that occur in degraded conditions. Further research is necessary to understand the extent to which visual speech (particularly visual form and timing information) produced by native and non-native talkers are differed and to what extent such differences affect speech perception in noise.

5. REFERENCES

[1] Bertolino, P. 2012. Sensarea: An authoring tool to create accurate clickable videos. In 10th workshop on content-based multimedia indexing (pp. 1–4). Annecy, France.

[2] Boersma, P., Weenink, D. 2010. Praat: Doing phonetics by computer (Version 5.1.32) [Computer program]. Retrieved 30 April, 2014. Available from http://www.praat.org/

[3] Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A. A. 2009. The Natural Statistics of Audiovisual Speech. *PLoS Comput Biol, 5*, e1000436.

[4] Davis, C., Kim, J. 2006. Audio-visual speech perception off the top of the head. *Cognition, 100*, B21-B31.

[5] Flege, J.E. 1995. Second language speech learning: Theory, findings, and problems. In Strange, W. (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research*. Timonium MD: York Press, 233-277.

[6] Flege, J. E., Bohn, O.-S., Jang, S. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25*, 437-470.

[7] Greenberg, S., Carvey, H., Hitchcock, L., Chang, S. 2003. Temporal properties of spontaneous speech - a syllable-centric perspective. *J. Phonetics, 31*, 465-485.

[8] Kawase, S., Hannah, B., Wang, Y. 2012. Effects of visual speech information on native listener judgments of L2 speech intelligibility and accent. Pronunciation in Second Language Learning and Teaching conference (PSLLT), Vancouver, Canada, August 24-25.

[9] Kawase, S., Hannah, B., Wang, Y. 2014. The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers. *J. Acoust. Soc. Am., 136*, 1352-136.

[10] Kim, J., Davis, C. 2014. How visual timing and form information affect speech and non-speech processing. *Brain and Language*, 137, 86-90.

[11] Macleod, A., Summerfield, Q. 1990. A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *Brit. J. Audiol., 24*, 29-43.

[12] McGurk, H., MacDonald, J. 1978. Visual influences on speech perception processes. *Percept. Psychophys., 24*, 253-257.

[13] Munro, M. J. 1998. The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition, 20*, 139-154.

[14] Nielsen, K. 2004. Segmental differences in the visual contribution to speech intelligibility. *J. Acoust. Soc. Am*., *115*, 2606.
[15] Sommers, M. S., Tye-Murray, N., Spehar, B. 2005. Auditory-Visual Speech Perception and Auditory-Visual Enhancement in Normal-Hearing Younger and Older Adults. *Ear and Hearing*, 26, 263-275.
[16] Sumby, W. H., Pollack, I. 1954. Visual Contribution to Speech Intelligibility in Noise. *J. Acoust. Soc. Am*., *26*, 212-215.
[17] Wang, Y., Behne, D. M., Jiang, H. 2009. Influence of native language phonetic system on audio-visual speech perception. *J. Phonetics*. 37, 344-356.
[18] 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.
[19] Wilson, I., Gick, B. 2014. Bilinguals Use Language-Specific Articulatory Settings. *J. Speech Hear. Res.*, *57*, 361-373.