NON-NATIVE DISCRIMINATION ACROSS SPEAKING STYLE, MODALITY, AND PHONETIC FEATURE

Sarah E. Fenwick[1,2], Catherine T. Best[1,3], and Michael D. Tyler[1,2]

[1]MARCS Institute, [2]School of Social Sciences and Psychology, [3]School of Humanities and Communication Arts, University of Western Sydney, Sydney, Australia
s.fenwick@uws.edu.au; c.best@uws.edu.au; m.tyler@uws.edu.au

ABSTRACT

Discriminating between certain non-native contrasts can be difficult. The Perceptual Assimilation Model [1] predicts that when two non-native phones are assimilated to the same native language category, as equally good or poor versions, discrimination should be poor (a single-category assimilation). However, it is not known to what extent visual and/or clearly articulated speech might assist cross-language speech perception. Monolingual Australian English listeners discriminated two single-category Sindhi consonant contrasts (/ʈ/-/ʈ̪/, /ɓ/-/ɓ/), across auditory-only (AO) and auditory-visual (AV) conditions, in clear and citation speech. For /ɓ/-/ɓ/ (a laryngeal feature difference), AV contrasts were discriminated more accurately than AO contrasts in citation speech, but not in clear speech, while for /ʈ/-/ʈ̪/ (a place-of-articulation difference) there was AV benefit for clear, but not for citation speech. These results highlight that while perceivers attempt to utilize even subtle gestural differences, speaking style and modality differentially contribute to the success of discriminating across non-native contrasts.

Keywords: speaking style, modality, cross-language speech perception, discrimination.

1. INTRODUCTION

The Perceptual Assimilation Model (PAM) [1] addresses the relative ease or difficulty of discriminating non-native speech contrasts by monolingual listeners. Although support for PAM has developed from auditory-only (AO) research, its discrimination predictions apply equally to auditory-visual (AV) speech, as according to PAM the perceiver directly perceives *articulatory gestures*, rather than solely the acoustic signal. According to PAM, discrimination accuracy is dependent on the way in which pairs of non-native phones are assimilated to native language categories. For example, when both non-native phones in a contrast are assimilated to two different native categories (a two-category assimilation), discrimination is predicted to be excellent. Alternatively, two non-native phones may be assimilated to the same native language category. If there is a perceived goodness-of-fit difference, where one phone is perceived as a better exemplar of the native language category than the other (a category-goodness assimilation), then discrimination is predicted to range from good to very good. But, if there is no reported goodness-of-fit difference, then discrimination is predicted to be poor (a single-category assimilation). Non-native listeners find it difficult to discriminate single-category contrasts because they have learned to tune out those phonetic differences in their native language.

In many cases where native speech segments are difficult to discriminate in AO conditions, they may be easier to distinguish visually [4, 13, 15]. Gaining a perceptual advantage with the addition of visual speech to audio speech is known as AV benefit, and it is useful in *adverse* listening conditions, such as in foreign language listening, when the listener has imperfect phonological knowledge of the L2 [8, 12]. For example, in a study investigating the most effective L2 training method for Japanese and Korean learners of English, AV benefit was found not only in the identification post-test of the perceptually difficult non-native contrast /r/-/l/, but pre-test results also revealed higher identification scores for the AV condition, as opposed to the AO condition [10]. This result demonstrates that even before perceptual training, the addition of visual speech can be beneficial.

The degree of visual information available, as well as speaking style used, has been shown to aid the perception of non-native speech. [11] examined the ability of L2 learners to make use of visual information that differed in degree of visual informativeness, as a function of consonant place of articulation /p/, /b/ and /v/ (labial/labiodental consonants) and manner of articulation /r/ and /l/. It was found that non-native participants had more AV benefit when identifying contrasts that differed in the more visually salient place of articulation differences than the less visually salient differences between the two liquid consonants. Similarly, speaking more clearly assists listeners' comprehension in difficult communicative situations (e.g., [7]). This type of speaking style has been labeled clear speech, and is generally characterized as a slow, exaggerated style of production [3, 16]. The current study compared

discrimination performance across clear and *citation* speech, which is speech that is used in experimental settings, and is not inherently naturalistic, nor for the purposes of the listener (e.g., [5, 14]). This comparison of clear and citation speech, across stimuli with a varying degree of available visual information, will allow an assessment of the generalizability of previous cross-language speech perception theories, such as PAM [1], which have focused on testing AO citation speech, and it will also extend what is known about clear speech in cross-language research to the auditory-visual domain.

Although it may be said that clear speech enhances perception, and the addition of visual speech enhances perception, the combined effect that AV speech and speaking style have on the discrimination of difficult non-native contrasts is largely unknown. Given that the predicted discrimination accuracy of two-category and category-goodness contrasts is too high for the present goal, as it ranges from good to excellent, and the factors discussed thus far should be most beneficial in 'difficult' listening conditions, this paper will focus on single-category contrasts. Therefore, the aim of the current study is to examine the relative discrimination performance of single-category contrasts, across clear and citation speaking styles, in both AO and AV conditions. Sindhi has been chosen as the target stimulus language because it has a large number of contrasting phones not found in English that only differ on POA or a less visually distinct laryngeal difference. It is predicted that there will be a significant interaction between modality and feature difference, such that AV speech should be discriminated more accurately than AO speech, but only when the contrast varies by POA, and is therefore visually distinct. Moreover, speaking style should significantly interact with feature difference and modality, such that this AV advantage should be more pronounced under clear speech conditions.

2. METHOD

2.1. Participants

Twenty monolingual native Australian English (AusE) speakers (12 females, 8 males, M_{age} = 21.4, age range = 18-38 years) were recruited from the first year psychology pool at the University of Western Sydney, in return for course credit.

2.2. Stimuli

AV speech recordings were conducted in a sound dampened booth at the MARCS Institute, University of Western Sydney. The selected speaker was a native Sindhi 35-year-old female, from Radhan, Pakistan. The full set of Sindhi consonants was produced in /Ca/ nonsense syllables, in citation and clear speaking styles. Each utterance was recorded in front of a black back-drop with a two-point lighting arrangement (Studio-Lite Photon Beard, Highlight 110), with diffusion paper, via a Sony-HXR HD camera (NX30p) at 50fps (1280 x 720), and a RØDE shotgun microphone (NTG-3; 44.1 kHz sampling rate via a MOTU ultra lite MK3 sound card) was positioned approximately 30 cm away from the speaker's mouth. To ensure there were no substantial head positioning differences between utterances, a concealed non-restrictive headrest was used.

To elicit citation speech the speaker was instructed to read the syllable that appeared on a computer screen. However, for the clear-speech trials, the speaker was provided with the additional instruction to "say each syllable clearly as if you were communicating with someone who had difficulty understanding what you were saying" [7]. Therefore, the speaking style productions differed in that for citation speech, the speaker was simply reading aloud, whereas for clear speech, the speaker was provided the knowledge that their speech was to assist a listener's intelligibility. To prepare the stimuli, the audio recording was high-pass filtered (70 Hz), and the onset, offset and duration of each utterance was determined using Praat [2]. Based on these acoustic measurements, the command-line tool FFmpeg [6] was used to extract the video-only segments. All videos began with the speaker's mouth in a closed, neutral position, and the vowel portion of each token was truncated to 75 ms, in order to reduce the possibility of discrimination judgements based on differences in acoustic vowel information [9, 17]. In some cases, the truncated duration of the vowel was lengthened (<10 ms) to ensure that the video counterpart of a token was not segmented in the middle of a frame, and a cosine off-ramp was applied to the last 5 ms, to prevent auditory clipping. All audio and corresponding video segments were synchronised using Adobe Premiere Pro, and a window was manually positioned over each video stimulus, such that only the speaker's mouth and throat were visible.

2.2.1 Stimulus Selection

The final tokens were selected from among those that a native Sindhi speaker correctly identified 100% of the time. PAM [1] assimilation types were then determined by means of an AO citation categorization task, in which 35 monolingual AusE listeners categorized all Sindhi AO citation stimuli to AusE phonological categories. Only citation AO tokens were presented in order to select stimuli based on typical PAM [1] testing conditions. This means that

any differences in discrimination across conditions may be relative to the predictably low AO performance. Two single-category contrasts were selected: the /ʈ/-/ʈ̪/ contrast that only differs by a POA constriction, and the /b/-/ɓ/ contrast, which differs on a less visible laryngeal difference. The laryngeal voiced bilabial stop /b/ and implosive bilabial stop /ɓ/ contrast was consistently categorized as English 'b' 98% and 96% of the time, respectively, while the POA voiceless retroflex stop /ʈ/ and the voiceless dental stop /ʈ̪/ were both consistently categorized as English 'd' 76%, and 87% of the time, respectively. For both contrasts, there were no significant differences between goodness-of-fit ratings. The participants were also presented with two two-category contrasts, /f/-/v/ and /ɓ/-/ɖ̪/, and two category-goodness contrasts; /ʈ̪/-/d̪/ and /d̪/-/ɖ/, but in the interests of brevity we will focus here on the results for the single-category contrasts only.

2.3. Procedure

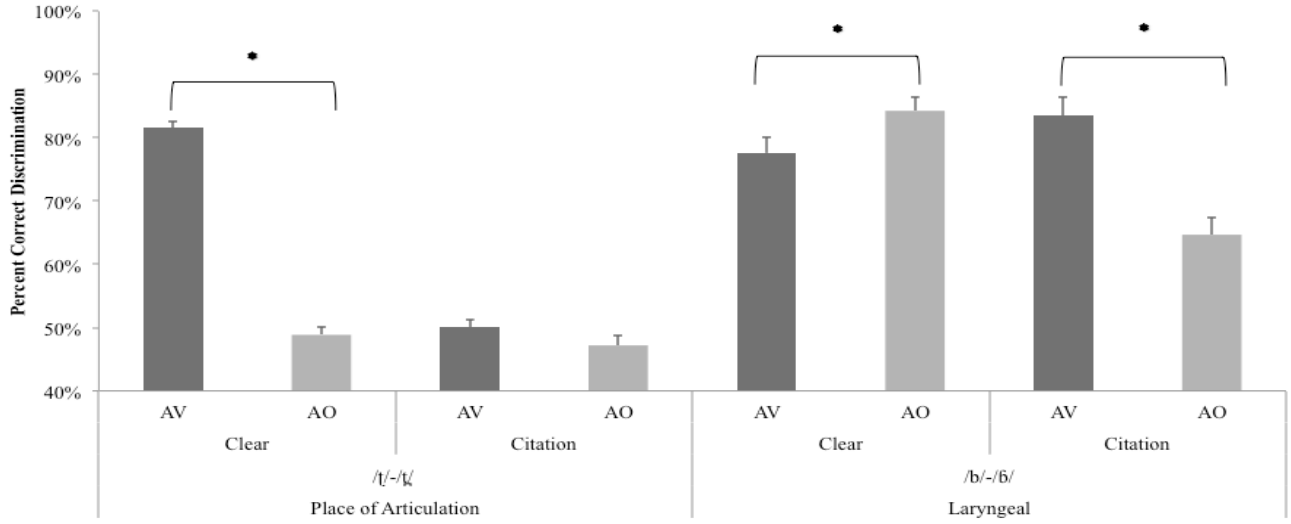
Each participant completed two 1.5 hr testing sessions on the same day. All contrasts were presented in four counterbalanced AXB discrimination conditions; AO citation, AO clear, AV citation and AV clear. On each trial of an AXB discrimination task, participants were presented with three stimulus tokens, separated by a 1-s inter-stimulus interval. They were asked to indicate whether the consonant of the second token (X) matched that of the first (A) or third (B), by pressing '1' or '3' on the computer keyboard. Once a response had been registered for a trial, the following trial began after 1s. If a response had not been registered within 3.5s, the trial was repeated at random.

Each contrast was presented in separate blocks of 48 trials, presented in random order. There were four possible trial types (i.e., AAB, BAA, BBA, ABB), presented 12 times each. To avoid acoustically based discrimination judgments, the consonants that belonged to the same phonological category were physically different tokens.

3. RESULTS

A 2 (Modality: AV, AO) x 2 (Speaking Style: clear, citation) x 2 (Feature Difference: POA, laryngeal) within-subjects analysis of variance (ANOVA) was conducted on the mean percent correct discrimination scores. There was a main effect of speaking style, $F(1, 19) = 39.37$, $p < .001$, $\eta_p^2 = .67$, with clear speech (M = 73%) discriminated more accurately than citation (M = 61%), a main effect of modality, $F(1, 19) = 72.44$, $p < .001$, $\eta_p^2 = .83$, with AV contrasts (M = 73%) discriminated more accurately than AO (M = 61%), and a main effect of feature difference, $F(1, 19) = 92.46$, $p < .001$, $\eta_p^2 = .79$, with the contrast that differed by a laryngeal feature (M = 78%) discriminated more accurately than the POA contrast (M = 57%). A significant 2-way interaction was found between speaking style and feature difference, $F(1, 19) = 13.61$, $p = .002$, $\eta_p^2 = .02$, as well as between feature difference and modality, $F(1, 19) = 15.25$, $p = .001$, $\eta_p^2 = .45$. Finally, as can be seen in Figure 1, simple effects analyses with a Bonferroni correction revealed differences within a significant 3-way interaction between feature difference, speaking style and modality, $F(1, 19) = 53.67$, $p < .001$, $\eta_p^2 = .74$.

Figure 1: *Percent correct discrimination results across auditory-only (AO) and auditory-visual (AV) conditions, in clear and citation speech for contrasts with a place-of-articulation versus laryngeal feature difference. An asterisk indicates a statistically significant difference between group means (p < .05), and error bars represent standard error of the mean.*



For the POA contrast, AV speech only assists discrimination when the participant is presented with clear speech, $F(1, 19) = 46.55$, $p < .001$, $\eta_p^2 = .71$, but not citation speech, $F(1, 19) = 3.50$, $p = .08$. For the laryngeal contrast, the reverse was found, where the addition of visual speech only assisted discrimination performance when presented in citation conditions, $F(1, 19) = 29.78$, $p < .001$, $\eta_p^2 = .61$, and for this same contrast, AO discrimination improved under clear speech conditions, in comparison to AO citation speech conditions, $F(1, 19) = 30.34$, $p < .001$, $\eta_p^2 = .62$, but there was no additional benefit of AV presentation. In fact, the discrimination of clear speech was more accurate for AO than AV presentation, $F(1, 19) = 6.81$, $p = .02$, $\eta_p^2 = .26$.

4. DISCUSSION

The aim of this study was to examine whether AV speech and clear speech improve discrimination for difficult non-native contrasts that vary in the amount of available visual information. As predicted, the addition of visual speech was most beneficial for the POA contrast in clear speech. Articulation of /t̪/-/t̪̲/ in clear speech appears to have provided additional information that perceivers can make use of in the visual, but not the auditory modality.

Interestingly, there was no AV benefit for the POA contrast in the citation condition, and an AV benefit *was* found for the laryngeal contrast in citation speech, which should provide the least amount of visual information. This may be explained by examining the visual information available in the two speaking styles.

Figure 2: *Visual representation of laryngeal /b/-/ɓ/ contrast, across citation and clear speech.*

Although /b/-/ɓ/ is characterized by a laryngeal feature difference, there also appears to be a difference between the degree of lip compression in citation speech for /b/ versus /ɓ/, which is most pronounced in the citation speaking style (see Figure 2).

These results therefore suggest that perceivers will take advantage of all sources of information available. Examined for the first time within AV cross-language conditions, clear speech differentially contributes to the likelihood of successfully discriminating non-native contrasts. The speaker's efforts to enhance the /b/-/ɓ/ distinction in clear speech, which improved AO discrimination for non-native perceivers, appears to have reduced the visual distinctiveness and prevented any AV benefit. Therefore, clear speech may not always visually enhance contrasting phonetic information.

As previously stated, PAM [1] predicts that single-category contrasts should be discriminated poorly, however the current study has shown that this is not necessarily the case when participants are presented with visual and clearly articulated non-native speech. To determine whether the PAM predictions are upheld across modalities, future PAM-oriented research may shed light on the above-average single-category discrimination. In this experiment we selected stimuli based on AO citation assimilation patterns, however examining the assimilation of AV and clear speech may demonstrate a shift in assimilation type, due to the addition of supplementary gestural information. For instance, in AV clear speech conditions, participants may detect a goodness-of-fit difference between the phones in the POA /t̪/-/t̪̲/ contrast, such that the assimilation type shifts to a category-goodness assimilation, thus explaining the good to very good discrimination patterns observed in this study.

Overall, the present study confirms that visual speech can enhance the perception of difficult non-native speech contrasts, however, discrimination performance is modulated by the contrast and speaking style presented. Additionally, the relatively few AV cross-language speech perception studies available generally focus on second-language learners, but as we have shown here, AV speech may still be beneficial with naïve perceivers, who have had no prior exposure to the target stimulus language. Finally, phonetic features alone cannot predict the degree of AV benefit, such that even subtle differences between laryngeal contrasts, such as lip compression, may be used to the perceiver's advantage. Accordingly, future research focusing on the effect of native language attunement on speech perception should compare the assimilation patterns for AO and AV speech, and compare the influence of speaking style on native versus non-native speech perception.

5. REFERENCES

[1] Best, C. T. 1995. A direct realist view of cross-language speech perception. In: Strange, W. (ed), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Baltimore: York Press, 171-206.

[2] Boersma, P., Weenink, D. 2014. Praat: Doing phonetics by computer. Retrieved from http://www.praat.org/

[3] Bradlow, A. R., Bent, T. 2002. The clear speech effect for non-native listeners. *J. Acoust. Soc. Am.* 112, 272-284. doi:10.1121/1.1487837

[4] Chen, T. 2001. Audiovisual speech processing. *Signal Processing Magazine, IEEE* 18, 9-21.

[5] Face, T. L. 2003. Intonation in Spanish declaratives. *Catalan J. of Ling.* 2, 115-131.

[6] FFmpeg. 2014. Retrieved from https://www.ffmpeg.org/

[7] Gagné, J. P., Rochette, A. J., Charest, M. 2002. Auditory, visual and audiovisual clear speech. *Speech Communication* 37, 213-230.

[8] Grant, K. W., Seitz, P. F. 1998. Measures of auditory-visual integration in nonsense syllables and sentences. *J. Acoust. Soc. Am.* 104, 2438-2450.

[9] Guion, S. G., Flege, J. E., Akahane-Yamada, R., Pruitt, J. C. 2000. An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *J. Acoust. Soc. Am.* 107, 2711–2724. doi: 10.1121/1.428657

[10] Hardison, D. M. 2003. Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics* 24, 495–522. doi:10.1017/S0142716403000250

[11] Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., Chung, H. 2006. The use of visual cues in the perception of non-native consonant contrasts. *J. Acoust. Soc. Am.* 119, 1740-1751. doi:10.1121/1.2166611

[12] Lecumberri, M. L. G., Cooke, M., Cutler, A. 2010. Non-native speech perception in adverse conditions: A review. *Speech Communication* 52, 864–886. doi:10.1016/j.specom.2010.08.014

[13] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264, 746–748.

[14] Peterson, G. E., Barney, H. L. 1952. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24, 175-184.

[15] Rosenblum, L. D., Saldaña, H. M. 1996. An audiovisual test of kinematic primitives for visual speech perception. *J. Exp. Psy.: Hum. Percep. & Perform.* 22, 318–331.

[16] Smiljanić, R., Bradlow, A. R. 2009. Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Lang. & Ling. Compass* 3, 236–264. doi:10.1111/j.1749818X.2008.00112.x.

[17] Tyler, M. D., Fenwick, S. E. 2012. Perceptual assimilation of Arabic voiceless fricatives by English monolinguals. *Proc. Interspeech 2012, Portland, USA*, 911-914.