A COMPARISON OF AUDIOVISUAL AND AUDITORY-ONLY TRAINING ON THE PERCEPTION OF SPECTRALLY-DISTORTED SPEECH

Najwa Alghamdi, Steve Maddock, Guy J. Brown and Jon Barker

Department of Computer Science, University of Sheffield
{amalghamdi1,s.maddock,g.j.brown,j.p.barker}@sheffield.ac.uk

ABSTRACT

Recent research suggests that using visual speech in auditory training can improve auditory-only speech perception. The long term aim of our work is to investigate this approach for hearing-impaired users, in particular cochlear-implant users. In the pilot study presented in this paper, we use spectrally-distorted speech to train two different groups of normal hearing subjects: native English and non-native, English-speaking Saudi listeners. Our pilot study suggests that both groups attain similar improvement in audio-only speech perception when visual speech is introduced into the training process. This may provide evidence that cochlear implant users would benefit from introducing visual speech in training, given that the reduced processing abilities of non-native listeners for native speech could be compared to the reduced processing abilities of cochlear implant users as a result of the inherent noise in the processing of sound by a cochlear implant.

Keywords: Speech perception; CI Simulation; Auditory training; Visual Speech; Cochlear Implant

1. INTRODUCTION

The use of technology such as cochlear implants (CI) has brought dramatic changes to users' lives. Nonetheless, many challenges still face CI users, which prevent them from utilising the full potential of their assistive technology, in particular during speech perception in noise. Solutions to these speech perception challenges can be sought by using audiovisual training approaches that make use of "brain plasticity" and the use of vision in speech perception.

The human brain has the ability to adapt after a long period of hearing deprivation, an ability known as brain plasticity, which can be induced by learning and behavioural changes. The central auditory system (CAS) that is responsible for speech perception can be restructured by auditory training to enhance the perceptual experience of CI users [18]. The auditory perceptual learning achieved by such auditory training has an influential impact on the CAS's response to known and novel auditory stimuli, and hence enhances listening ability. For example, researchers have found that perceptual learning gained from speech-in-noise auditory training can significantly enhance speech perception in noise for normal hearing, hearing aid users and CI users [4, 16].

The role of vision in speech perception has been extensively researched. Auditory signals that are difficult to hear, are usually easier to see [19]. CI users that relied on vision during speech perception in their previous period of deafness, show a high capacity for audiovisual synergy compared with normal hearing people [15]. CI practitioners have exploited this to offer audiovisual training (AV) to CI users to learn communication strategies such as speech production and speech/lip-reading.

Recent evidence has found a link between introducing visual speech in auditory training (AV training) that aims to improve auditory skills, and inducing CAS plasticity. The perceptual learning gained from AV training was found to be more effective in enhancing CI-simulated speech perception by normal hearing listeners than audio-only (AO) training [3, 14, 10]. In AO training, auditory signals guide top-down perceptual learning. However, when auditory signals are compromised by noise, such as in the case of CI users, the available visual signals offered by the AV training can provide external support to help develop auditory perceptual learning [3]. This can consequently shape a perception experience that can be later utilised by listeners to comprehend novel stimuli even in an auditory-only situation.

This paper builds on the previous work in AV training to enhance hearing in AO situations. As with previous work, we use spectrally-distorted speech to simulate how speech is processed by a CI user (although this does not necessarily reflect the hearing experience of CI users, which can be worse than the simulation [6]). We also use normal hearing listeners. The difference in our paper is that we use two different normal hearing groups: native

and non-native listeners.

If we consider non-native speakers as analogous to CI users, in that they both deal with adversity in perception, our pilot study may thus provide some evidence that CI users would benefit from similar training. We can consider a number of factors to support the comparison between non-native speakers and CI users. First, non-native listeners tend to perform worse in speech identification in adverse conditions compared with native listeners. During speech in noise perception, non-native listeners face two problems: auditory signal degradation and linguistic knowledge [17, 11]. The acoustic variability of non-native speech, as well as the phonemic and phonological rules of the listener's first language, pose challenges in vowel and consonant contrast and sentence perception [9, 8]. Moreover, both cochlear implant noise and non-native linguistic knowledge can be classified as internal adverse conditions, and can cause a failure to map the acoustic/phonetic features to lexical units [1, 13].

2. METHOD

2.1. Subjects

Two groups of normal hearing participants were recruited in two different geographic locations: 9 native English listeners, and 12 non-native Saudi listeners (with IELTS score >= 5.5). All subjects were in the age range 18-35 (M = 24 years, SD = 4.5 years). The hearing ability of each subject was screened using a pure tone audiometric test. The Saudi participants were sub-grouped into equal groups, A_s and V_s, and the English participants were sub-grouped into groups of 4 and 5 participants, A_e and V_e, where A and V denote audio-only training and audio-visual training (using video of a talker's face), respectively.

2.2. Stimuli

The audiovisual GRID corpus [5] was used to provide training stimuli that featuring the same speaker, which consists of sentences such as "bin blue at A 2 please" with the following syntax [command:4] [colour:4] [preposition:4] [letter:25] [digit:10] [adverb:4], where the number of choices for each keyword is indicated in the square brackets (and letters=25, since W is not included [5]). The GRID audio was spectrally distorted using an eight-channel sine-wave vocoder (AngelSim[1]). It was hypothesized that normal hearing listeners can perform in a comparable way to CI users when hearing no more or less than 8 channels [7]. Consistent with [2], the following settings were used for the vocoder: a bandpass filter was first applied to divide the signal into 8 channels between 200 to 7,000Hz (slope=24dB/octave); each channel was then low-pass filtered by 160Hz (slope=24dB/octave) to obtain the envelope; the envelope of each channel modulated a sine wave that replaced the signal frequency. The GRID videos were processed using the FFMPEG [2] framework to replace the audio in each video with the vocoded one.

2.3. Procedures

First, all subjects performed an auditory-only (AO) pre-test of 12 trials to set a baseline level for each subgroup. After that, three training sessions were used. Each training session used a different set of 20 stimuli. Each set of 20 stimuli was made up of 10 vocoded stimuli repeated twice, but with the whole set of 20 randomised in order. The A_s and A_e subgroups received AO training and AO testing in each session, while the V_s and V_e subgroups received AV training and AO testing in each session. The AO testing at the end of each training session was used to track learning milestones for all subgroups. After completing all training sessions, all subjects performed an AO post-test of twelve trials in order to assess their training gain from the AO pre-test. As the GRID corpus was used to provide stimuli, the training/testing task for each subject was to identify the colour, the letter and the digit that corresponded to the played stimulus, and enter these using three button presses on a labelled keyboard. During training, after submitting their input for a stimulus, it was then replayed with added subtitles whether the input was correct or not. During testing, no such feedback was provided.

3. RESULTS AND DISCUSSION

3.1. Post-training Impact on Perceptual Learning

Figure 1 compares the pre- and post-training scores for the auditory only tests. The V (V_s and V_a) groups achieved a higher training gain than the A (A_s and A_e) groups. Within the V groups, the error bars show a significant difference between the Saudi and English groups before the training [T=2.26, P=0.04], which is possibly due to two "noise masks": the internal linguistic knowledge issues in the Saudi group and peripheral intelligibility degradation issues as a result of the vocoding process. The graph

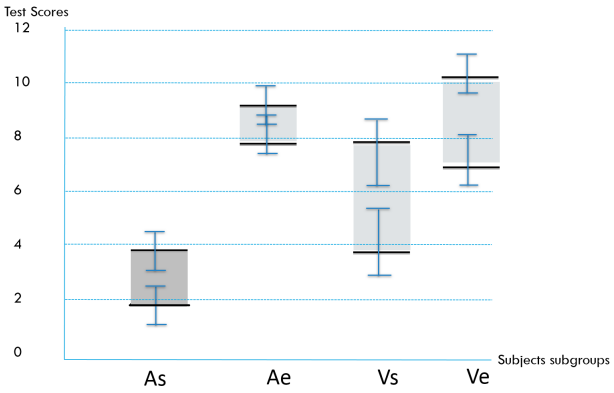

Figure 1: Overall identification results during audio-only test for Saudi participants, A_s and V_s, and English participants, A_e and V_e. Black horizontal lines represents audio-only test mean score before (bottom line) and after (top line) training for each subgroup. A = AO training and AO testing; V = AV training and AO testing; s = Saudi; e = English.

also shows that AV training increases the V_s group's intelligibility of the vocoded speech, reaching a comparable level to the pre-training native English group (A_e and V_e). This may indicate that the visual signal helped the $V_s$subjects to adapt to the vocoded speech, although it is unclear if this adaptation had enhanced the intelligibility of the vocoded speech from the internal or/and the peripheral noise masks mentioned above. Whilst the A_s subjects improved, they were still unable to reach a comparable level to the baseline levels of the English subgroups. This was confirmed by a T-test result that showed a significant difference in letter identification between A_s after the training and A_e before (P=0.01) and after (P=0.0004) the training, given the fact that letter identification is the most challenging task for all subjects due to the need to select from a larger set with high variance (25 letters) as opposed to colours (4) and digits (10).

One point to mention is that the V_s group performed better than the A_s group in the pretest, suggesting that they were a better group of listeners. This makes the post-test harder to interpret since the V_s group might have shown greater improvement because they started with better performance in the first place. It is clear that, given that a small data set is used, a larger study needs to be undertaken to fully test the observations.

3.2. During the Training

Figure 2 depicts the identification scores during the training for all subgroups. Within the Saudi group, we observed that the V_s subgroup outperformed

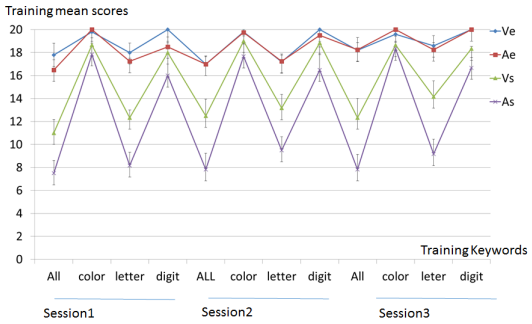

Figure 2: Mean identification results during training blocks for Saudi participants, A_s and V_s, and English participants, A_e and V_e. A = AO training and AO testing; V = AV training and AO testing; s = Saudi; e = English.

the A_s subgroup in the letter identification task. Within the English group, there was no significant difference between A_e and V_e subgroups for all keyword identification tasks. Between the Saudi and the English groups, there was no significant difference in colour and digit identification tasks scores during the training, but there was a significant difference in letter identification scores.

3.3. Letter Confusion Matrices

Confusion matrices (Figure 3) were derived to understand the possible sources of confusion the subjects had while identifying the letter keywords during the post AO test. Circle diameters represent the confusion mean rate for respective row-column pairs. The strong main diagonal pattern of circles in Figure 3, V_e and A_e, clearly illustrate that the English subgroups were less confused than the Saudi subgroups.

For the English subgroups, no significant difference was spotted between V_e and A_e. For the Saudi subgroups, V_s showed better overall performance in letter identification as illustrated by the diameter of the circles on the main diagonal. This was confirmed by a T-test results that showed the significant difference in post AO test mean scores [T=2.63, P=0.013] between V_s and A_s. The V_s group achieved higher scores in the identification of letters that are constructed from diphthongs [a, e, i, u] with a significant difference [T=3.12, P=0.02]. Since vowels differ in the frequency of the first formants (F1 and F2), and given that F1 and F2 are correlated with jaw height and tongue position [12], visual signals may contribute to enhance the intelligibility of diphthongs by the V_s group, while the A_s group showed more compressed vowel space with high confusion between A and E (confusion mean=31%) and I and O (confusion

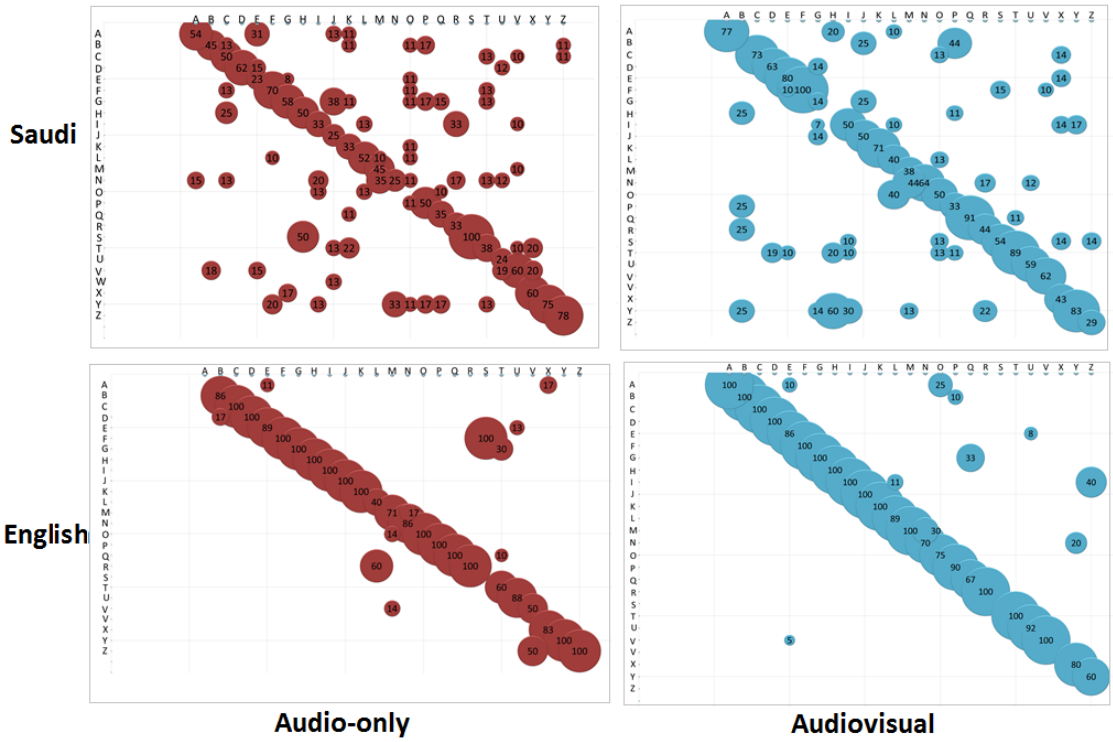

Figure 3: Letter Confusion Matrices. Top : A_s and V_s , Bottom: A_e and V_e

mean=13%). The V_s group also outperformed A_s with a significant difference [P=0.002] in identifying the nasal sound in N and voiceless plosive sounds.

On the other hand, the V_s group showed confusion between visually similar letters, for example (G and D) and (P and B), compared with the A_s group. In these letters, visual signals may have impeded learning the invisible sounds (such as postalveolar and velar) that were bounded by visible ones (such as vowels and alveolar), for example, the invisible sound /ʒ/ in G /dʒiː/.

As the voicing information is generally affected by the vocoding process, all subjects reported difficulty in discriminating between voiced B and voiceless P. The Saudi subjects were more affected due to a language specific factor, given that the Saudi Arabian dialect's inventory lacks the voiceless sound P. However, in the V_s group, the confusion rate was higher (confusion mean=44%), indicating that visual cues may have impeded the learning of voicing discrimination of the pair B and P.

4. CONCLUSIONS

We have reported a pilot study that investigates the impact of introducing visual speech in auditory training that is aimed at enhancing auditory only speech perception. Spectrally-distorted speech stimuli were used to simulate how a CI processes perceived speech. In contrast to other research work, we recruited both native and non-native listeners, considering the hearing experience of non-native listeners for spectrally-distorted speech as analogous to CI users. The post-test results for this small pilot study suggest that the AV training groups (V_s and V_e) achieved the highest training gain compared with the AO groups(A_s and A_e). However, a larger study is required to confirm this initial observation.

The experiment also gave some insight into how the vocoding process can affect the transfer of place, manner and voicing information. Although visual signals seemed to slightly improve the transfer of some nasality and voicing information as reflected by the letter confusion matrices, they may impede learning the discrimination of visually similar sounds. The next step in our work will be a larger study followed by a similar experiment on a group of CI and hearing aid users.

Acknowledgement: This research has been supported by the Saudi Ministry of Education, King Saud University, and the European Community 7th Framework Programme Marie Curie ITN INSPIRE (Investigating Speech Processing in Realistic Environments).

5. REFERENCES

[1] Assmann, P., Summerfield, Q. 2004. The perception of speech under adverse conditions. In: *Speech processing in the auditory system*. Springer 231–308.

[2] Bent, T., Buchwald, A., Pisoni, D. B. 2009. Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *The Journal of the Acoustical Society of America* 126(5), 2660–2669.

[3] Bernstein, L. E., Auer Jr, E. T., Eberhardt, S. P., Jiang, J. 2013. Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Frontiers in neuroscience* 7.

[4] de Boer, J., Thornton, A. R. D. 2008. Neural correlates of perceptual learning in the auditory brainstem: efferent activity predicts and reflects improvement at a speech-in-noise discrimination task. *The Journal of Neuroscience* 28(19), 4929–4937.

[5] Cooke, M., Barker, J., Cunningham, S., Shao, X. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120(5), 2421–2424.

[6] Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., McGettigan, C. 2005. Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General* 134(2), 222.

[7] Dorman, M. F., Loizou, P. C. 1998. The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels. *Ear and hearing* 19(2), 162–166.

[8] Højen, A., Flege, J. E. 2006. Early learners discrimination of second-language vowels. *The Journal of the Acoustical Society of America* 119(5), 3072–3084.

[9] Ji, C., Galvin, J. J., Chang, Y.-p., Xu, A., Fu, Q.-J. 2014. Perception of speech produced by native and nonnative talkers by listeners with normal hearing and listeners with cochlear implants. *Journal of Speech, Language, and Hearing Research* 57(2), 532–554.

[10] Kawase, T., Sakamoto, S., Hori, Y., Maki, A., Suzuki, Y., Kobayashi, T. 2009. Bimodal audio–visual training enhances auditory adaptation process. *Neuroreport* 20(14), 1231–1234.

[11] Lecumberri, M. L. G., Cooke, M., Cutler, A. 2010. Non-native speech perception in adverse conditions: A review. *Speech Communication* 52(11), 864–886.

[12] Löfqvist, A., Sahlén, B., Ibertsson, T. 2010. Vowel spaces in swedish adolescents with cochlear implants. *The Journal of the Acoustical Society of America* 128(5), 3064–3069.

[13] Mattys, S. L., Davis, M. H., Bradlow, A. R., Scott, S. K. 2012. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes* 27(7-8), 953–978.

[14] Pilling, M., Thomas, S. December 2011. Audiovisual cues and perceptual learning of spectrally distorted speech. *Language and Speech* 54(4), 487–497.

[15] Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., Barone, P. 2007. Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences* 104(17), 7295–7300.

[16] Sweetow, R. W., Sabes, J. H. 2007. Technologic advances in aural rehabilitation: Applications and innovative methods of service delivery. *Trends in Amplification* 11(2), 101–111.

[17] Tabri, D., Chacra, K. M. S. A., Pring, T. 2011. Speech perception in noise by monolingual, bilingual and trilingual listeners. *International Journal of Language & Communication Disorders* 46(4), 411–422.

[18] Tremblay, K., Kraus, N., McGee, T. 1998. The time course of auditory perceptual learning: neurophysiological changes during speech-sound training. *Neuroreport* 9(16), 3557–3560.

[19] Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., Jones, C. J. 1977. Effects of training on the visual recognition of consonants. *Journal of Speech, Language, and Hearing Research* 20(1), 130–145.

[1] http://www.tigerspeech.com/angelsim/angelsim_about.html

[2] https://www.ffmpeg.org/