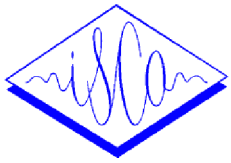
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 - September 3, 2007

Benefits of facial and textual information in understanding of vocoded speech.

Sharon M. Thomas, Michael Pilling

MRC Institute of Hearing Research, University Park, Nottingham, NG9 2RD, U.K.

st@ihr.mrc.ac.uk

Abstract

Exposure to audiovisually presented vocoded speech is more effective than exposure to auditory-only vocoded speech in improving the subsequent ability to understand vocoded speech [1]. In addition, improvements in the audiovisual training condition were more rapid and greater in magnitude than in the auditory-only condition. The current study was conducted to establish whether exposure to concurrent textual information also results in improvements in the ability to recognize vocoded speech. Baseline measures of identification performance with auditory-only vocoded speech sentences were assessed for 45 participants. Participants then performed a speech identification task, where they were exposed to vocoded speech in either audiovisual (Group 1), auditory-only (Group 2), or auditory-only with concurrent text conditions (Group 3). Following exposure, participants were tested again on identification performance with auditory-only vocoded speech. Exposure to concurrent text improved subsequent understanding of vocoded speech, to a level similar to that seen with audiovisual speech exposure. In a second experiment, groups of normal hearing adults were exposed to vocoded non-lexical nonsense words in auditory-only with text and audiovisual speech presentation conditions. Exposure to both nonsense audiovisual and concurrent text conditions improved subsequent understanding of lexical auditory-only vocoded speech, and there was no difference between the levels of improvement. In summary, to computer-based audiovisual and concurrent text exposure improves the ability to recognize vocoded speech over exposure to auditory stimuli alone. This effect does not appear to be dependent on exposure to lexical items.

1. Introduction

Cochlear implants (CIs) are of enormous benefit in facilitating speech understanding in the severely and profoundly deaf. This is despite the fact that the overall quality of speech information transmitted by a CI is highly impoverished and heavily distorted. Though CIs are effective in the transmission of the overall amplitude envelope information of a speech signal, there is only minimal transmission of spectral information. Additionally, what spectral information is transmitted is shifted upwards in frequency because of tonotopic misalignment of electrode positions in the cochlea. Partly because of these problems, an extended period of several months of exposure and rehabilitation is typically required before CI users plateau in their speech recognition performance, though much of the gain is in the initial months post-implant [e.g. 2].

The principal characteristics of CI speech can be simulated by noise-vocoding and spectrally shifting a speech signal [3]. Research on the perceptual learning mechanisms underlying CI speech adaptation has mostly looked at how normal hearing listeners adapt to such spectrally-shifted vocoded speech and interpolated such findings to CI users [e.g. 3; 4; 5]. This research has demonstrated that normal hearing participants can show somewhat rapid adaptations towards this novel form of speech, with substantial gains in intelligibility found after only short periods of exposure. Rosen et al. [4] presented speech stimuli that simulated the consequences of 6.5mm tonotopic misalignment. After nine twenty-minute sessions of training using Connected Discourse Tracking, identification of words in sentences improved from less than 1% to 30%. Stacey and Summerfield [6] also investigated the effectiveness of a 2-alternative forced-choice training to improve the ability of normal-hearing subjects to understand speech that had been vocoded. Training produced an improvement of 9.4% in target word identification.

Cues from visual speech contribute much to the understanding of the speech signal. Particularly in noisy conditions, the benefits of seeing as well as hearing a talker are well established in normal hearing populations [e.g., 7; 8; 9; 10; 11; 12; 13]. Benefits are also seen even when the auditory signal is clear and unambiguous [14; 15]. A further indication of the power of visual speech influence on auditory perception occurs with the McGurk effect [16], where different auditory and visual signals combine to form a new percept that was not presented in either modality alone. A typical example is when an auditory syllable (e.g., /ba/) is presented in synchrony with a visual image of a talker articulating a different syllable (e.g., /ga/). Here, observers generally perceive a syllable (e.g., /da/) that was not presented either visually or auditorily but represents a combination of both sources of information.

One of the benefits of visual speech is that it provides salient cues that allow the perceiver to disambiguate confusable phonetic features, such as consonant place cues and vowel formant frequency [20]. These are cues present largely in the in the fine spectral structure of an acoustic signal, and thus impoverished in the CI. In unambiguously highlighting cross-modal relations between a talker's facial movements and the resultant sounds, perceivers might learn to explicitly distinguish the different articulatory features present in vocoded speech. Rosen, Faulkner, and Wilkinson, [4] provided evidence consistent with this possibility. They used audiovisual presentations with vocoded speech in early training blocks in a study looking at adaptation. They found the number of recognized words in sentences rose considerably after exposure to audio and audiovisual presentations of vocoded speech. However, because participants all undertook the same training it is not possible to assess the effect of visual speech itself of the adaptation process. Benefits of audiovisual over unimodal auditory speech perception have been found in CI users. Kaiser, Kirk, Lachs and Pisoni [21] examined how postlingually deafened adult CI users combine visual

information from visual speech with auditory cues in a word recognition task. Results showed that word recognition performance was higher when participants viewed audiovisually presented words, than when they listened to auditory-only presentations.

The potential of exposure to audiovisually presented vocoded speech to enable better subsequent understanding of vocoded auditory speech alone was explored by Thomas & Pilling, [1]. Their results showed that observing a talking face in addition to hearing an auditory signal enhances perception of vocoded speech. Exposure to an auditory speech identification task, using normal and vocoded speech, enabled participants to better subsequently identify words in vocoded speech sentences. However, exposure to audiovisually presented vocoded speech resulted in larger improvements than in either of the auditory only speech training conditions. In a further experiment, they demonstrated that audiovisual training resulted in more rapid and greater improvements than the auditory only training condition. These results suggest that intervention and training programs aiming to increase CI users spoken language understanding may wish to exploit the advantages conveyed by visual speech on speech processing. However, the source of the observed advantage for audiovisual speech over auditory speech alone should be explored further. More specifically, In order to explore fully the role that audiovisual speech exposure plays in enabling subsequent improved understanding of impoverished speech, it is important that this role is distinguishable from that played by cognitive top-down processing. There are several example of the powerful effect of top-down influences on speech perception. Remez, Rubin, Pisoni and Carrell, [23] demonstrated that when listeners are exposed to sine-wave speech, the content is perceived as unintelligible. However, if they are informed of the identity of the original sentence, then the sine-wave speech becomes subsequently comprehensible. Similar effects have been reported by Davis, Johnsrude, Hervais-Adelman, Taylor, and McGettigan, [5] using vocoded speech. These effects are interpreted as owing to the top-down knowledge enabling more efficient perceptual grouping of the distorted speech. As we have observed, one of the benefits of visual speech is that it provides salient cues for disambiguation of confusable phonetic features. However, it may be the case that higher-level processes provide the dominant influence when beneficial effects of exposure to audiovisual speech are observed. Specifically, the availability of visual speech as well as the auditory signal may simply have improved the intelligibility of the speech stream, thus providing cognitive feedback on auditory speech content. A direct comparison between the benefits afforded by audiovisual speech and those afforded by more explicit lexical information (e.g., audiotextual stimuli) for subsequent understanding of distorted speech will help to disambiguate the direct or indirect nature of visual speech influences.

The following experiment was designed to compare the effects of exposure to vocoded speech in audiovisual, auditory-only, and auditory with concurrent text on intelligibility gains in vocoded speech. If exposure to concurrent text with auditory vocoded speech does facilitate subsequent understanding of auditory-only vocoded speech, then participants should demonstrate greater intelligibility of vocoded speech after concurrent text is withdrawn, compared to those who have only been exposed to auditory-only speech.

2. Experiment 1

2.1. Method

2.1.1. Stimuli

Speech processing: Auditory speech processing was carried out to simulate the sound distortions associated with transmission through an 8-channel cochlear implant, with a 6mm basalward shift on the basilar membrane. The auditory processing to achieve cochlear implant simulation was carried out using the same parameters as Stacey & Summerfield [5]. Auditory speech was processed off-line using software routines written in the Matlab© programming environment. Prior to this processing, audio tracks were treated using Audacity© audio editing software to remove any adverse sounds unrelated to the speech and to suppress any background noise. The principles underlying cochlear implant simulation are described in greater detail in Shannon et al. [2] and Rosen et al [3]. Input speech was first passed through 8 elliptical IIR analysis filters (simulating the eight independent channels of the cochlear implant). Resultant waveforms from each filter were half-way rectified and low-pass filtered at 160 Hz. Envelopes were then multiplied by low-pass filtered white noise (10 Hz) and each filtered by an elliptical output filter. Output filters had their band edges increased in frequency by a value comparable to a 6 mm offset. Finally, the eight separate channels were summed together to produce a single speech signal.

Training and testing materials: Training and test materials consisted of the auditory, audiovisual and auditory with text recordings of 298 spoken BKB sentences [21]. Each sentence had 3 key words. The allocation of sentences to test sessions gave the same number of key words present in each test session. All materials were spoken by an adult male talker, recorded in a sound-deadened room in front of a neutral white background. Recordings were made using a Sony DSR200AP digital camcorder with the audio recorded through a Sony ECM-44B condenser microphone attached below the talker's face. All materials were recorded onto Sony DVCAM digital tape. The camcorder recorded at 25 frames per second. Audio was recorded at a sample rate of 48 kHz. Audiovisual recordings were spliced into separate video-clips using Final Cut Pro© video-editing software. A 1000 ms still-frame was put at the beginning and end of each clip. The auditory stream of all video-clips was processed to simulate a cochlear implant (see above). Finally, all video-clips were converted into Apple QuickTime format and processed to give a 500 ms fade in and fade out from black.

Text stimuli were displayed onscreen in Helvetica Neue font at 45 point. The text presentation started as closely as possible with the onset of the auditory speech, then scrolled across screen and disappeared 14cm to the right of centre. The text was presented at a speed matching as closely as possible that of the auditory speech.

2.1.2. Participants

45 normal-hearing adults took part in the study. All spoke English as their native language and had lived in Britain or Ireland for at least 10 years.

2.1.3. Procedure

Throughout the testing and exposure session, group 1 was presented with auditory-only vocoded speech sentences. Group 2 was presented with audiovisual vocoded speech sentences every second block. Group 3 was presented with audiovisual vocoded speech sentences with concurrent text every second block. The task was the same throughout, with participants required to type in to the computer as much of the sentence as they were able to decipher. The experimental session took approximately one hour to complete. Each session contained 76 BKB sentences, containing 235 keywords (some sentences had 3 keywords, others 4). All sentences were presented in a random order. No sentence was repeated. Sentences were scored using the 'loose keyword' method [21]. Participants were required to type in to the computer as much of the sentence as they were able to decipher. Guessing was encouraged. Where a participant was totally unsure of the speech in a particular test sentence they were instructed to type "I don't know". The tasks took approximately one hour to complete. The experiment comprised of one session comprising 57 blocks, each of 5 sentences. Test and training blocks were alternated, with the first and last block as test blocks.

2.2. Results

Figure 1 shows the percentages of words correctly identified in tests across exposure condition (auditory only vocoded speech, audiovisual vocoded speech, auditory vocoded speech with text).

ANOVA was applied to the test data, with condition as an independent factor with three levels corresponding to exposure condition (audiovisual, auditory-only vocoded speech, and auditory vocoded speech with text). There was an overall main effect of training condition; $F(1,2)=4.36$, $p<.05$. Post hoc tests (Newman-Keuls) revealed that both the audiovisual and auditory-only with text conditions differed from auditory-only ($ps<0.05$). However, there was no significant difference between audiovisual and auditory with text conditions ($p=1$).

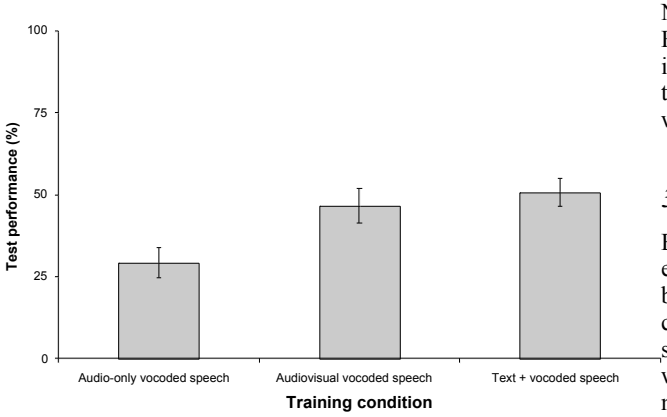


Figure 1: Overall group test performance (%, with SE) following exposure to the three different experimental conditions.

2.3. Discussion

Experiment 1 demonstrated that exposure to vocoded speech in auditory-only, audiovisual, and auditory with text conditions effectively improved subsequent understanding of auditory-only vocoded speech. Exposure to audiovisual speech and auditory vocoded speech with text was equally effective for the subsequent understanding of auditory-only vocoded speech.

These results suggest that the improvements gained from exposure to audiovisual speech owe more to the ability of audiovisual speech to access high-level linguistic content, rather than any clarification afforded by visual speech per se. Specifically, we could conclude that higher-level processes are the dominant influence on listeners' segmentation of distorted speech. If this is the case, then exposure to non-lexical audiovisual speech, or non-lexical auditory speech with text should not result in improvements in subsequent ability to understand vocoded speech. We examined this possibility in Experiment 2.

3. Experiment 2

3.1. Method

Experiment 2 sought to examine whether exposure to vocoded nonsense speech in audiovisual and audiotext conditions leads to improvements in the subsequent ability to recognise auditory-only vocoded speech. To explore this question, groups of normal hearing adults were trained using audiovisual vocoded nonsense speech and auditory only vocoded nonsense speech with concurrent nonsense text.

3.1.1. Participants.

28 normal-hearing adults took part in the study. All spoke English as their native language and had lived in Britain or Ireland for at least 10 years. None had participated in Experiment 1.

3.1.2. Stimuli

Nonsense sentences were based on the syllabic structure of BKB sentences used in the exposure trials of Experiment 1. Nonsense sentences contained no words found in English. For instance, "he's sucking his thumb" was transformed into "burs gilan zon reeb". "The boy has black hair" was transformed into "ja vix ku mor tul". Nonsense sentences were spoken by the same talker as used for Experiment 1.

3.1.3. Procedure

Before exposure, participants were presented with two examples of nonsense sentences. These were not vocoded but presented in clear speech. The exposure session consisted of a total of 105 vocoded nonsense speech sentences presented either audiovisually or auditory only with concurrent nonsense text. Participants were tested on normal (lexical), vocoded BKB sentences immediately after exposure. The task was the same throughout, with participants required to type in to the computer as much of the sentence as they were able to decipher. All other aspects of Experiment 2 were the same as for Experiment 1.The experimental session took approximately one hour to complete.

3.2. Results

Figure 2 shows the percentages of vocoded auditory-only words correctly identified after exposure across condition (auditory only vocoded nonsense speech exposure with text, and audiovisual vocoded nonsense speech exposure).

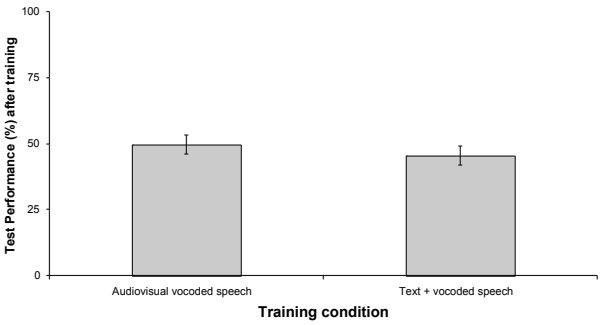


Figure 2: Overall group test performance (%, with SE) following exposure to the two experimental conditions.

Exposure to both audiovisual nonsense vocoded speech and auditory-only nonsense vocoded speech with text improved subsequent understanding of auditory-only vocoded speech. There was no difference between the levels of improvement (p=.33).

3.3. Discussion

Experiment 2 demonstrated that exposure to nonsense sentences in audiovisual and audio plus text presentation conditions subsequently enabled participants to better identify real words in vocoded speech sentences. These results indicate that it is not necessary for the content of vocoded speech to be lexical in order for beneficial effects of exposure to be observed.

4. General Discussion

The results confirm that computer-based exposure to vocoded speech in auditory-only and audiovisual conditions is effective in improving the ability of naïve listeners to understand spectrally distorted speech. Audiovisual exposure is again more effective than auditory-only, and the current study has also demonstrated auditory exposure with concurrent text is as effective as audiovisual exposure. This implies that the source of the observed advantage for audiovisual speech over auditory speech alone is not exclusive to visual speech, but extends to other disambiguating information, in this case, text. However, this disambiguation does not rely on the lexicality of the stimulus sentences. Exposure to nonsense vocoded speech with visual and textual information also improved subsequent perception of lexical vocoded auditory speech. Given the close relationship between speech production and facial movement, it is surprising that text (which does not enjoy this direct relationship with speech) appears to be as effective as visual speech in enabling adaptation to vocoded speech. Given that the effects do not seem to be based on lexicality, further studies are required to establish what it is that visual and textual information share in that enables the disambiguation of spectrally-distorted speech. For now, we can conclude that computer-based audiovisual or audio with text exposure may prove more beneficial for cochlear-implant users than auditory exposure alone. Further work is needed to establish what aspects of the visual stimulus and the task conditions are necessary for the effect to occur.

5. Acknowledgements

Thanks to Sam Irving for audiovisual recordings, and Paula Stacey for advice on auditory stimulus construction.

6. References

[1] Thomas, S.M., and Pilling, M. (2007). Benefits of audiovisual exposure using a simulation of a cochlear-implant system. Under revision. Ear & hearing.

[2] Clark, G.M. (2002). Learning to hear and the cochlear implant. In Fahle, M., and Poggio, T., (eds.) Perceptual learning: MIT Press: 147-160.

[3] Shannon, R.V., Zeng, F.G., and Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues, Journal of the Acoustical Society of America, 104, 2467-2476.

[4] Rosen, S., Faulkner, A., & Wilkinson, L. 1999. "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants." J. Acoust. Soc. Am., 106 (6), 3629-36.

[5] Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Tayler, K., & McGettingan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence From the Comprehension of Noise-Vocoded Sentences. Journal of Experimental Psychology:General, 134, 222-241.

[6] Stacy, P., & Summerfield, A.Q. (2005) Auditory-perceptual training using simulation of a cochlear-implant system: A controlled study. Proceedings of ISA Workshop on Plasticity in Speech Perception (PSP2005), London, UK.

[7] Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–15.

[8] Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech & Hearing Research, 12, 423-425.

[9] Erber, N.P. (1975). Auditory-visual perception of speech. Journal of Speech & Hearing Disorders 40: 481-492.

[10] MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. British Journal of Audiology, 12,131-141.

[11] MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. British Journal of Audiology, 24, 29-43.

[12] Middleweerd, M.J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. Journal of the Acoustical Society of America, 82, 2145-2147.

[13] Walden, B.E., Grant, K.W., & Cord, M.T. (2001). Effects of amplification and speechreading on consonant recognition by persons with impaired hearing. Ear Hear. 22: 333-341.

[14] Reisberg, D., McLean, J, & Goldfield, A (1987). Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli. In B. Dodd &

R. Campbell (Eds.), Hearing by eye: The psychology of lip-reading (pp. 97-113). Hillsdale, NJ: Erlbaum.

[15] Arnold, P. & Hill, F. (2001). Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. British Journal of Psychology, 92, 339–355.

[16] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

[17] Green, K.P. 1997. The use of auditory and visual information during phonetic processing: implications for theories of speech perception. In B. Dodd & R. Campbell (Eds.), Hearing by eye: The psychology of lip-reading (pp. 97-113). Lon.: Erlbaum.

[18] van Wassenhove, Grant, K.W., & Poeppel, D. 2005. Visual speech speeds up the neural processing of auditory speech. PNAS, 102, 1181-6

[19] Calvert. G. & Campbell. M. 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr. Biol. 10:649-57

[20] Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (eds.) Hearing by eye: The psychology of lip reading Hillsdale, New Jersey: Lawrence Erlbaum Associates.

[21] Kaiser, A., Kirk, K., Lachs, L., P. & Pisoni (2003). Talker and Lexical Effects on Audiovisual Word Recognition by Adults With Cochlear Implants. Journal of Speech, Language, and Hearing Research, 46, 390-404.

[22] Bench, J., & Bamford, J. (Eds.). (1979). Speech-hearing Tests and the Spoken Language of Hearing-impaired Children. London: Academic Press.

[23] Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212* (4497), 947-949.