Hearing Research 336 (2016) 17–28

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research paper

The contribution of visual information to the perception of speech in noise with and without informative temporal fine structure

Paula C. Stacey [a, *], Pádraig T. Kitterick [b], Saffron D. Morris [c], Christian J. Sumner [c]

[a] Division of Psychology, Nottingham Trent University, Burton Street, Nottingham NG1 4BU, UK
[b] NIHR Nottingham Hearing Biomedical Research Unit, Ropewalk House, 113 The Ropewalk, Nottingham NG1 5DU, UK
[c] MRC Institute of Hearing Research, University Park, Nottingham NG7 2RD, UK

ARTICLE INFO

Article history:
Received 3 December 2015
Received in revised form
6 April 2016
Accepted 11 April 2016
Available online 13 April 2016

Keywords:
Audio-visual
Visual speech
Temporal fine structure
Sine-wave vocoding
Cochlear implants

ABSTRACT

Understanding what is said in demanding listening situations is assisted greatly by looking at the face of a talker. Previous studies have observed that normal-hearing listeners can benefit from this visual information when a talker's voice is presented in background noise. These benefits have also been observed in quiet listening conditions in cochlear-implant users, whose device does not convey the informative temporal fine structure cues in speech, and when normal-hearing individuals listen to speech processed to remove these informative temporal fine structure cues. The current study (1) characterised the benefits of visual information when listening in background noise; and (2) used sine-wave vocoding to compare the size of the visual benefit when speech is presented with or without informative temporal fine structure. The accuracy with which normal-hearing individuals reported words in spoken sentences was assessed across three experiments. The availability of visual information and informative temporal fine structure cues was varied within and across the experiments. The results showed that visual benefit was observed using open- and closed-set tests of speech perception. The size of the benefit increased when informative temporal fine structure cues were removed. This finding suggests that visual information may play an important role in the ability of cochlear-implant users to understand speech in many everyday situations. Models of audio-visual integration were able to account for the additional benefit of visual information when speech was degraded and suggested that auditory and visual information was being integrated in a similar way in all conditions. The modelling results were consistent with the notion that audio-visual benefit is derived from the optimal combination of auditory and visual sensory cues.

1. Introduction

Speech perception in normal-hearing listeners is very resilient to distortions in the auditory signal and the presence of background noise. In contrast, understanding speech in background noise is difficult for adults with hearing impairment (Davis, 1989; Kramer et al., 1998) and is particularly problematic for users of cochlear implants (CI) whose device degrades the spectral and temporal information in speech (Schafer and Thibodeau, 2004; Wolfe et al., 2009; Fu et al., 1998; Skinner et al., 1994). Shannon et al. (1995) showed that when signals were presented in quiet, listeners with normal hearing were able to tolerate a dramatic reduction in the amount of spectral and temporal information present in the speech signal before there was any appreciable effect on performance. The 'noise-vocoding' technique used by Shannon et al. (1995) involved: (1) dividing the speech signal into a limited number of frequency bands; (2) extracting the slow amplitude modulations or 'temporal envelope' within each frequency band; and (3) using these envelopes to modulate a wide-band random-noise carrier signal which was then filtered by the same filters used in stage (1). The use of a random-noise carrier has the effect of replacing the informative high-rate fluctuations in frequency near the centre-frequency of each band with non-informative fine structure. As the first two stages of this process mimic the processing stages implemented by a speech processor of a cochlear implant, vocoders have been widely used to investigate the difficulties experienced by users of cochlear implants.

* Corresponding author.
E-mail addresses: paula.stacey@ntu.ac.uk (P.C. Stacey), padraig.kitterick@ntu.ac.uk (P.T. Kitterick), saffron90@live.co.uk (S.D. Morris), chris@ihr.mrc.ac.uk (C.J. Sumner).

http://dx.doi.org/10.1016/j.heares.2016.04.002

The inability of cochlear implants to convey informative temporal fine structure cues has severe consequences for the ability of cochlear-implant users to perceive speech in the presence of background noise (e.g. Schafer and Thibodeau, 2004), and this difficulty has been replicated using noise-vocoding in normally-hearing individuals (Qin and Oxenham, 2003; Ihlefeld et al., 2010; Rosen et al., 2013). Qin and Oxenham (2003) investigated speech perception in noise with 4-, 8-, and 24-channel vocoders. Normal-hearing listeners were presented with IEEE sentences, and the signal-to-noise ratio (SNR) at which performance was 50% correct (known as the Speech Reception Threshold, SRT_{50}) was estimated by varying the relative levels of speech and noise. When speech was unprocessed and presented in single-talker background noise, participants could achieve 50% correct performance at an SNR of −10.3 dB. When speech was then processed by an 8-channel vocoder, listeners required the level of the speech to be 6.4-dB higher than the noise to reach the same performance level. The addition of more spectral channels improved performance with the vocoder but a positive SNR (+0.7 dB) was still required to report 50% of keywords correctly even in the 24 channel condition. Qin and Oxenham (2003) concluded that the reduction of pitch cues found in the temporal fine structure and low frequency harmonics of speech may be responsible for this performance detriment. Somewhat lower levels of susceptibility to the presence of noise have been reported for speech processed using a 'sine-wave vocoder' in which the informative temporal fine structure is replaced with sine waves rather than noise (Whitmal et al., 2007). There is some evidence that sine-wave vocoders match the percept of cochlear-implant users more closely than noise-band vocoders (e.g. Dorman et al., 1997) and are better at preserving the envelope fluctuations present in speech (e.g. Whitmal et al., 2007; Dau et al., 1999).

Although the impact of removing informative temporal fine structure cues has been studied extensively for audio-only situations, its impact on the audio-*visual* perception of speech in noisy conditions has received little attention, despite this being the more ecologically relevant problem. Sumby and Pollack's (1954) seminal work with normal-hearing adults showed that word recognition improved considerably under audio-visual conditions compared to listening to the audio alone. In fact, the addition of visual speech information was found to be equivalent to increasing the signal-to-noise ratio by +15 dB compared with audio-only presentation. It is perhaps not surprising therefore that people with impaired hearing and users of cochlear implants gain considerable benefit from being able to see the faces of talkers (Erber, 1975; Kaiser et al., 2003; Tyler et al., 1997).

Kaiser et al. (2003) tested audio-only, visual-only, and audio-visual recognition of monosyllabic English words in both normal-hearing listeners and cochlear-implant users. Normal-hearing listeners were presented with words at −5 dB SNR, and cochlear-implant users were presented with words in quiet. The results showed that both groups of listeners performed best in the audio-visual condition in which word recognition scores were similar in both groups. There was some evidence that cochlear-implant users made better use of visual information when listening conditions were more difficult, such as when they were required to identify lexically difficult words (low frequency words with many phonetic neighbours, Luce and Pisoni, 1998). More recent studies have added support to the idea that people with cochlear implants may be better at integrating auditory and visual information than normal-hearing listeners (Rouger et al., 2007; Desai et al., 2008).

A number of previous studies have found that benefits from visual speech information depend on the nature of the auditory signal. Grant et al. (1985, 1991, 1994) investigated the way in which different sorts of degraded speech signals combined with visual speech cues. More recently, McGettigan et al. (2012) demonstrated greater benefits from visual speech information for speech lacking in auditory clarity, such that visual speech information boosted performance more for 2- and 4-channel noise-vocoded speech than it did for 6-channel vocoded speech.

These studies lead logically to the idea that the value of any sensory input is not fixed, but can depend of the value or nature of another sensory input; i.e. the visual signal is of greater value when the auditory input is degraded. This is consistent with the 'Principle of Inverse Effectiveness' (Lakatos et al., 2007; Tye-Murray et al., 2010) which asserts that the value of one modality will increase as the value of another declines. A number of models have been proposed to try to explain the nature of multisensory integration (Massaro, 1987; Blamey et al., 1989; Braida, 1991; Grant et al., 1998; Kong and Carlyon, 2007; Rouger et al., 2007; Micheyl and Oxenham, 2012). Models can be broadly categorised as to whether information is integrated in some raw sensory form before any decision is made ('pre-labelling') or after decision processes are applied separately to each modality ('post-labelling'; Braida, 1991; Peelle and Sommers, 2015).

Recently, Micheyl and Oxenham (2012) proposed a pre-labelling model based on Signal Detection Theory (SDT) to explain the capacity of normal-hearing listeners to integrate vocoded information in one ear with low-frequency acoustic information in the other ear. Their model and those applied in other similar studies suggested that the benefits of integrating electric and acoustic information can be explained as an additive interaction (Seldran et al., 2011; Micheyl and Oxenham, 2012; Rader et al., 2015) of the raw sensory information prior to any decision. Rouger et al. (2007) applied a post-labelling model to examine the properties of audio-visual integration, which assumes that decisions are made about individual cues prior to integrating these to make an overall decision. Their model is an extension of the 'probability summation model' (Treisman, 1998), which states that the probability of answering correctly is equal to the probability that either one or both of the modalities presented individually would result in the correct answer. Interestingly, Rouger et al.'s implementation of this model on their data suggested that integration across modalities operated differently in cochlear implantees and normal hearing subjects listening to noise-vocoded speech.

The current project systematically investigates the perception of sine-wave vocoded speech (labelled as ENV speech) at a range of SNRs, and compares this with performance in 'clear' speech conditions where informative temporal fine structure cues remain (labelled as TFS speech). The primary question of interest is whether the size of the benefit received from visual speech information depends on the presence of informative temporal fine structure information. This question was addressed using both open-set and closed-set tests of speech perception as we might expect to find differences between different types of speech tests (see Lunner et al., 2012). Not only were we interested in whether any numeric improvement in performance with the addition of visual information depended on the presence of TFS, but also whether any observed differences implied a difference in the underlying integration process. Three experiments are presented below; in the first participants completed an open-set sentence test using a between participants design, the second reports an open-set sentence test using a mixed participants design, and the third reports a closed-set sentence test using a mixed participants design. Background noise consisted of multi-talker babble. In each experiment we expected to find that visual speech information contributed more to understanding vocoded speech in background noise than to understanding clear speech in background noise. These results were interpreted within the framework of a SDT model.