The Impact of Automatic Exaggeration of the Visual Articulatory Features of a Talker on the Intelligibility of Spectrally Distorted Speech

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

Department of Computer Science

University of Sheffield

Abstract

Visual speech information plays a key role in supporting speech perception, especially when acoustic features are distorted or inaccessible. Recent research suggests that for spectrally distorted speech, the use of visual speech in auditory training improves not only subjects' audiovisual speech recognition, but also their subsequent *auditory-only* speech recognition. Visual speech cues, however, can be affected by a number of facial visual signals that vary across talkers, such as lip emphasis and speaking style. In a previous study, we enhanced the visual speech videos used in perception training by automatically tracking and colouring a talker's lips. This improved the subjects' audiovisual and subsequent auditory speech recognition compared with those who were trained via unmodified videos or audio-only methods. In this paper, we report on two issues related to automatic exaggeration of the movement of the lips/ mouth area. First, we investigate subjects' ability to adapt to the conflict between the articulation energy in the visual signals and the vocal effort in the acoustic signals (since the acoustic signals remained unexaggerated). Second, we have examined whether or not

this visual exaggeration can improve the subjects' performance of auditory and audiovisual speech recognition when used in perception training. To test this concept, we used spectrally distorted speech to train groups of listeners using four different training regimes: (1) audio only, (2) audiovisual, (3) audiovisual visually exaggerated, and (4) audiovisual visually exaggerated and lip-coloured. We used spectrally distorted speech (cochlear-implant-simulated speech) because the longer-term aim of our work is to employ these concepts in a training system for cochlear-implant (CI) users.

The results suggest that after exposure to visually exaggerated speech, listeners had the ability to adapt alongside the conflicting audiovisual signals. In addition, subjects trained with enhanced visual cues (regimes 3 and 4) achieved better audiovisual recognition for a number of phoneme classes than those who were trained with unmodified visual speech (regime 2). There was no evidence of an improvement in the subsequent audio-only listening skills, however. The subjects' adaptation to the conflicting audiovisual signals may have slowed down auditory perceptual learning, and impeded the ability of the visual speech to improve the training gains.

***Highlights*.**

1. An automatic method that exaggerates the visual articulatory features of talkers in audiovisual speech is presented.
2. An audiovisual conflict after-effect was found, but it becomes negligible after exposing subjects to exaggerated speech.
3. Exaggeration of the visual speech improves the audiovisual recognition of a number of phoneme classes.
4. Audiovisual recalibration to visually exaggerated speech may have impeded learning when used in the audiovisual training.

Keywords: Audiovisual training; Cochlear-implant simulation; Visual-speech enhancement; Lombard speech

1. Introduction

The robustness of human speech perception arises from a listener's ability to integrate and evaluate information from multiple sources. 'Audiovisual integration' refers to a listener's ability to utilise auditory and visual speech information in order to interpret the perceived message from the talker [1, 2]. The illusion of perceiving a new audio signal when listeners are presented with an incongruent audiovisual signal – known as the McGurk effect [3] - provides compelling evidence of the synergy of audio and visual speech during perception. This complementary support is also evident in adverse listening conditions, such as when listening to speech in noise, where visual speech cues can improve the speech intelligibility by 5–22 dB [4–7]. The external articulators (lips, teeth, and tongue) can provide a significant proportion of the overall visual speech information gathered from the face [8, 9]; McGrath [9] found that the accuracy of identifying monophthongal vowels reached 56% by lipreading the external articulators only. By visualising these external articulators' kinematic information using a point-light technique, Rosenblum *et al.* [10] found that such visualisation can substantially enhance the intelligibility of speech in noise. Compared with other facial movements that provide temporal cues only, kinematic information from the mouth can provide both speech information and temporal cues [11].

Visual speech also significantly contributes to speech intelligibility when listeners undergo 'internal' adverse conditions (i.e. when listeners suffer limitations in perceptual skills), such as in the case of cochlear implant (CI) users. CI users experience the noise inherited from CI-processed speech, since such speech lacks important spectral and temporal cues necessary for pitch perception [12–15]. In addition, the amount of acoustic information CI users can

*I am corresponding author

Email addresses: amalghamdi1@sheffield.ac.uk (Najwa Alghamdi*), s.maddock@sheffield.ac.uk (Steve Maddock), j.p.barker@sheffield.ac.uk (Jon Barker), g.j.brown@sheffield.ac.uk (Guy J. Brown)

receive is a function of various physical and physiological factors, including the number of implanted and activated electrodes and the severity of damage to the hearing nerve [16]. Fortunately, a talker's face provides a medium in which CI users can speech-read visibly distinguished phoneme categories that are difficult to hear, such as labial (anterior consonants like bilabials) or non-labial (posterior consonants) phonemes.

Attending to visual speech in face-to-face communication is not the only strategy that CI users utilise as a way of overcoming degraded acoustic cues. Auditory perception training (or *auditory training*), for example, can significantly improve CI-listening abilities in audio-only modalities. Such training aims to enhance the central auditory system (CAS) responses to sound (i.e. CAS plasticity) by using auditory perceptual learning [17]. Recent studies have found a link between induced CAS plasticity and the introduction of visual speech in auditory training (i.e. *audiovisual training*) in which auditory skills are improved. Such studies found that the perceptual learning gained from audiovisual training was more effective in enhancing CI-simulated speech perception by normal-hearing listeners than was the case with auditory training [18–21]. In auditory training, acoustic signals guide 'top-down' perceptual learning. When acoustic signals are compromised by noise, however (such as in the case of CI users), the available visual signals that the audiovisual training offers can provide external support to help develop auditory perceptual learning [18]. This can consequently shape a perceptual experience that listeners can later utilise to comprehend novel stimuli, even in auditory-only situations.

The benefits of visual speech may be affected by a range of talker-dependent factors, including lip emphasis [22–24], teeth and tongue visibility [2, 9, 25], facial hair [26], speaking style [24, 27], and talker's gender [28]. This raises the question of whether or not enhancing visual speech quality can increase the benefit of such speech. In a previous study [29], we found that enhancing the appearance of a talker's lips within audiovisual training stimuli can help to improve the audiovisual and subsequent auditory recognition of CI-simulated speech by non-native, normal-hearing subjects. Using image processing and computer graphics techniques, we simulated the talker wearing lipstick in the training stimuli. According to Lander and Capek's [22] observations of talkers who wore real lipstick, this is an effect that can improve lip-reading. We found improved training gains in audiovisual and audio-only sentence-recognition rates when the subjects were trained by audiovisual speech produced by a talker wearing simulated lipstick compared with the original unmodified audiovisual speech and the audio-only speech of the same talker [29].

Speaking style is a determining factor in the quality of visual speech [24, 27]. According to Lindblom's hypo-hyper (H&H) theory of speech production [31], talkers make articulatory energy modifications from hypo- to hyper-articulated speech in order to adapt to the demands of the listening situation. This may create a variety of speaking styles that exert different energy magnitudes in order to move the external articulators (Figure 1) [30]. This concept led us to investigate the transition from hypo- to hyper-articulated speech as a modification method for enhancing visual speech. Theobald *et al.* [32] addressed speaking-style exaggeration in 2D videos in order to support the forensic lip-reading of surveillance videos; they exaggerated the lip movements of a talker in a video by amplifying the principle components (PCs) of the talker's mouth shapes and appearance that were elicited from the video. They found

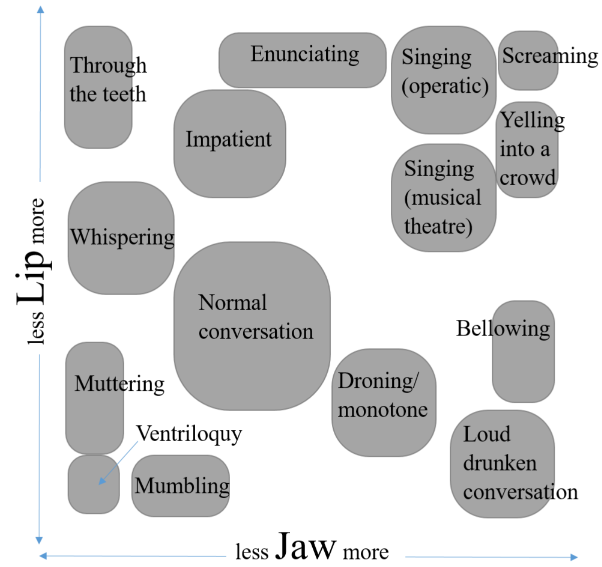

Figure 1: *The transition from hypo- to hyper-articulation in the Jaw Lip viseme model (JALI), which simulates visual speech by aggregating functions related to jaw motion and lip motion [30]: This shows that distinct speaking styles exert different articulatory-energy modifications. Permission to use this image has been granted by Edwards [30].*

improved lip-reading performance among inexperienced lip-readers.

The study we present in this paper investigates the impact of exaggerating the visual articulatory features (in particular the mouth kinematics) on the visual benefit of audiovisual speech. The main challenge of such an enhancement is that the combination of exaggerated visual signals and the original unexaggerated acoustic signals could create conflicting, incongruent inputs. The impact of exposure to conflicting inputs has been widely investigated in the behavioural-studies literature [3, 33–36]. These impacts can be classified as immediate biases or after-effects [33]. An example of immediate biases may be observed in spatial conflict (such as the ventriloquism effect), where visual stimuli can influence sound localisation [34], and in identity conflict (i.e. the McGurk effect) [3]. Another study also found that exposure to mismatched inputs can create an after-effect on perceptual modalities used for adapting to this conflict. For example, visual speech has the ability to recalibrate auditory perception after exposure to the conflicting audiovisual stimuli observed in the McGurk effect [33]. Audiovisual recalibration has also been observed in temporal audiovisual conflicts (such as in live broadcasts) to help adapt to time lags [37, 38].

Given these factors, our study has a twofold aim: first, to investigate the after-effects of listeners' exposure to the conflict between the articulation energy in the visual signal and the vocal effort in the acoustic signal (as well as determining whether the listeners will acquire the ability to adapt in order to overcome the audiovisual mismatch); and second, to study the impact of such modifications to audiovisual speech on improvements to audiovisual and subsequent auditory speech recognition when used in audiovisual training. To achieve these goals, a similar exaggeration method to Theobald *et al.* [32] was used in this study to model mouth shape variation. We exaggerated mouth motions by extrapolating the PCs of a talker's mouth shapes in a given video; we then applied 2D image warping to reanimate the video using the new exaggerated mouth motion. 2D image warping, which involves the geometric transformations that define a relationship between two images' pixels [39], is a well-known technique for facial modifications that is used for visualising plastic-surgery outcomes and in various entertainment platforms [40–42].

The audio part of all stimuli that were used in the experiment was spectrally distorted to simulate what a CI user might hear. (We should note, however, that CI simulation does not necessarily reflect the hearing experience of an actual CI user, whose hearing may be worse than the simulation, since CI users' inner ears in general are partially or fully damaged [43]). In order to consider the impact of the internal adverse condition caused by the damaged inner ear among CI users, we chose non-native listeners (where by non-native we mean non-UK-native. The listeners were female Saudi nationals - see Section 2.1 for full details) for our experiment. Previous studies have shown that non-native listeners' perception is degraded when listening to native speech in adverse conditions compared with native listeners. During speech-in-noise perception, non-native listeners face two problems: auditory signal degradation (which is analogous to the CI-simulated speech used in this experiment) and limited linguistic knowledge [44, 45]. Limited linguistic knowledge is classified as a form of internal adverse condition that can cause a failure to map acoustic/phonetic features to lexical units [46, 47]. There may, however, be some differences between native and non-native lipreaders in using visual cues [48]. The non-native lipreaders' linguistic experience of the native language [49] is an important factor that impacts the weighting of visual cues by the lipreaders. To control this variable among the subjects, we used the IELTS [International English Language Testing System] test score (in particular the listening band aspect) as a measure to select subjects. To avoid different English abilities influencing the result, subjects were automatically assigned into subgroups based on their pre-test scores (explained in Section 2.4) in order to create balanced subgroups in terms of their pre-test abilities. We may therefore consider our non-native subjects as being analogous to CI users, since both types of listeners cope with internal and external adversity in their perception of CI-processed speech.

In order to create a homogeneous subject group, we chose female subjects for our experiment. Behavioural studies have suggested gender differences in audiovisual perception: females have been found to be more sensitive to visual speech than males and are better speech-readers [50]. The audio-only and audiovisual stimuli of the training were taken from a selected talker in the audiovisual 'Grid' corpus [51]. Modifications to the audiovisual stimuli were applied by exaggerating the mouth movements and/or colouring the lips of the talker to create modified audiovisual stimuli. Individuals were given an audio-only pre-test and were trained in one of four alternative conditions: (1) audio only, (2) audiovisual, (3) exaggerated audiovisual, or (4) exaggerated audiovisual with simulated lipstick applied. The subjects were then tested again using audio-only stimuli. The method, results, and discussion of this study are presented in the following sections.

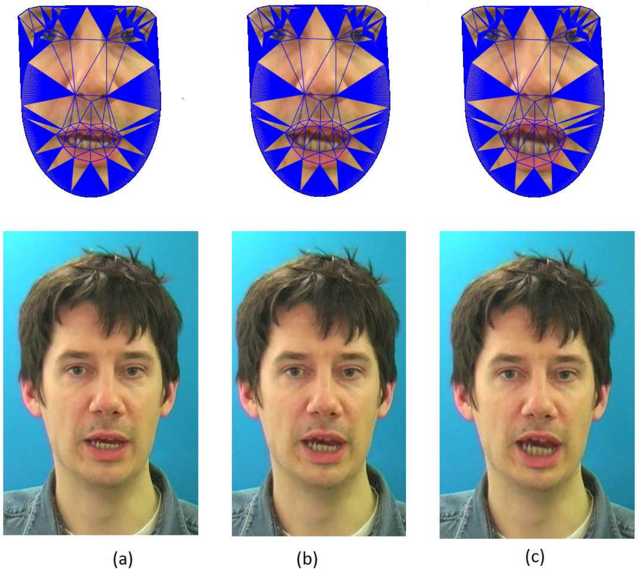

Figure 2: *Frame warping after estimating exaggerated mouth shapes: (a) the original frame; (b) and (c) frames under two levels of exaggeration effect (α = 1.5 and 2, respectively).*

2. Method

2.1. Subjects

The experiment was conducted at the female campus of King Saud University in Riyadh, Saudi Arabia. The subjects were 71 female non-native Saudi listeners (with minimum IELTS score = 5.5), each in the age range 18–40 years (mean = 24 years; standard deviation [SD] = 4.5 years). The hearing ability of each subject was screened using a pure-tone audiometric test. The participants were split into four groups: A (16 subjects), V (15 subjects), E (21 subjects), and $\text{E}_{lipstick}$ (19 subjects). See Section 2.3 for more details.

2.2. Training Stimuli

The audiovisual Grid corpus [51] was used to provide training and testing stimuli during the auditory-perception training. Each stimulus consisted 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 square brackets (and letters = 25, since 'W' was not included [51]). The subjects had to identify the colour, letter, and digit keywords in each stimulus. Using a single talker from the Grid corpus, four different modalities of stimuli were created:

1. Audio-only (A) stimuli, which were used for the A-group training as well as in the audio-only tests for all participants (see Section 2.3);
2. Audiovisual (V) stimuli, which were used in the V-group training;
3. Audiovisual with exaggeration effect (E) stimuli, which were used in the E-group training (one level of exaggeration was used, α =2); and

4. Audiovisual with exaggeration and lipstick effect ($\mathrm{E}_{lipstick}$) stimuli, which were used in the $\mathrm{E}_{lipstick}$-group training (Figure 5).

The implementation of the visual effects used in stimuli 3 and 4 is explained in Section 2.3. The audio-only Grid stimuli were created by spectrally distorting the Grid audio tracks using an eight-channel sine-wave vocoder (AngelSim[1]). All audiovisual stimuli (including those used in V, E, and $\mathrm{E}_{lipstick}$) were modified by replacing the original audio track of the Grid videos with the corresponding spectrally distorted (i.e. vocoded) audio-only Grid stimuli using the FFMPEG tool[2].

A set of 250 Grid sentences of the selected talker were randomly chosen to ensure that multiple subjects provided coverage of the same sentence. This set was then randomly split into ten sets of twelve sentences, ten sets of ten sentences, and five sets of six sentences. Four versions were produced for each of these sentences: an audio version, an audiovisual version, an exaggerated audiovisual version, and an exaggerated, lipstick-applied audiovisual version. Each subject was then assigned seven random sets of stimuli: two sets of twelve stimuli, three sets of ten stimuli, and two sets of six stimuli. The sets of twelve were used in audio-only (A) pre- and post-tests, while the three sets of ten and the two sets of six were used in the training process (see Section 2.4).

2.3. Exaggeration and Lipstick Effects

Selected audiovisual Grid videos (stimuli 2) were processed as a batch using Faceware Analyzer[3] tracking software, which is based on an active appearance model and uses the shape and texture information to elicit the facial landmarks of the talker. This process exports an XML file for each video that contains normalised (x,y) locations of facial landmarks in all video frames, including 30 landmarks of the face (eyebrows, eye corners, pupil, and nose) and 26 landmarks of the mouth (inner and outer lips). To correct for small variations over time in the talker-camera distance in a video, the mouth coordinates were normalised and translated to produce a zero-centred mouth space. The points were normalised by dividing the mouth landmarks in each frame k by a distance d_k, where d_k is the Euclidean distance between the midpoint of the inner corners of the eyes and the tip of the nose, since these are assumed to be unaffected by the articulation. To create the zero-centred mouth model space, the normalized mouth landmarks in frame k were translated by T to be aligned with the centre of the normalized mouth landmarks in the first frame, where T is formed from the 2D distance between the mouth centres. The k^{th} video frame may then be associated with two vectors: a mouth shape vector of 52 elements, expressed as:

$$Lip_k = \begin{bmatrix} x_1 & y_1 & \cdots & x_{26} & y_{26} \end{bmatrix}^T \tag{1}$$

and a face shape vector of 60 elements, expressed as:

$$Face_k = \begin{bmatrix} x_1 & y_1 & \cdots & x_{30} & y_{30} \end{bmatrix}^T \tag{2}$$

[1] http://www.tigerspeech.com

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

[3] http://facewaretech.com

A set of eigenvectors generated from the covariance matrix of a given training set can be used to approximate any of that set [32, 52]. Given that, we used a set of eigenvectors generated from the covariance matrix of mouth shapes from a given video $(Lip_1, Lip_2, \cdots Lip_n)^4$, to approximate any mouth shape, Lip_k, in that video as follows:

$$Lip_k \approx \overline{Lip} + Pb_k \tag{3}$$

where $\overline{Lip}$ is the mean mouth shape in the corresponding video, P is the matrix of t^5 eigenvectors with the highest eigenvalues (each column represents an eigenvector), expressed as:

$$P = \begin{bmatrix} p_{1,1} & p_{2,1} & \cdots & p_{t,1} \\ p_{1,2} & p_{2,2} & \cdots & p_{t,2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1,52} & p_{2,52} & \cdots & p_{t,52} \end{bmatrix} \tag{4}$$

and b_k is a t dimensional vector, expressed as:

$$b_k = \begin{bmatrix} b_{k,1} & b_{k,2} & \cdots & b_{k,t} \end{bmatrix}^T \tag{5}$$

and given by:

$$b_k = P^T(Lip_k - \overline{Lip}) \tag{6}$$

where b_k defines the contribution of each eigenvector in the representation of Lip_k, and can be seen as a measure of the distance between $\overline{Lip}$ and Lip_k [32]. Thus, multiplying b_k with a scalar $\alpha > 1$ can extrapolate (i.e. exaggerate) lip shape as follows:

$$newlip_k \approx \overline{Lip} + \alpha Pb_k \tag{7}$$

To project $newlip_k$ from the zero-centred model space back to its location in the video frame, $newlip_k$ is translated by T^{-1} then scaled by d_k. The exaggeration effect applied to the E and E$_{lipstick}$ stimuli used $\alpha = 2$. The video frames were then re-animated by applying a 2D piecewise linear warping method using the estimated exaggerated mouth shapes $newlip_k$ to apply the exaggeration effect (Figure 2). Each frame was partitioned using a triangulation algorithm given $newlip_k$, $Face_k$ and a ring that delimited the face (Figure 2: top row). The use of a ring restricts the exaggeration impact to the face only. Backwards linear transformation was then applied to each triangle, such that pixels in the target image are inversely mapped to the source image pixels and sampled based on that mapping [39]. Figure 3 shows the visual exaggeration effects on viseme (visual phoneme) classes extracted from videos of one speaker from the Grid dataset [53]; the first column represent the original viseme shape, while the second column represents the viseme after exaggeration by $\alpha = 2$.

In order to create a lipstick effect on the exaggerated video, a similar process as detailed in [29] and illustrated in Figure 4 was applied to the exaggerated videos (Figure 5).

[4] n ≈ 51 shapes; Each grid video contains 64 frames resulting in 64 mouth shapes, however, we excluded silence frames when calculating the covariance matrix (13 frames on average).

[5] t=10 eigenvectors can account for 90-99 % of the lip variance, based on tests made on selected videos.

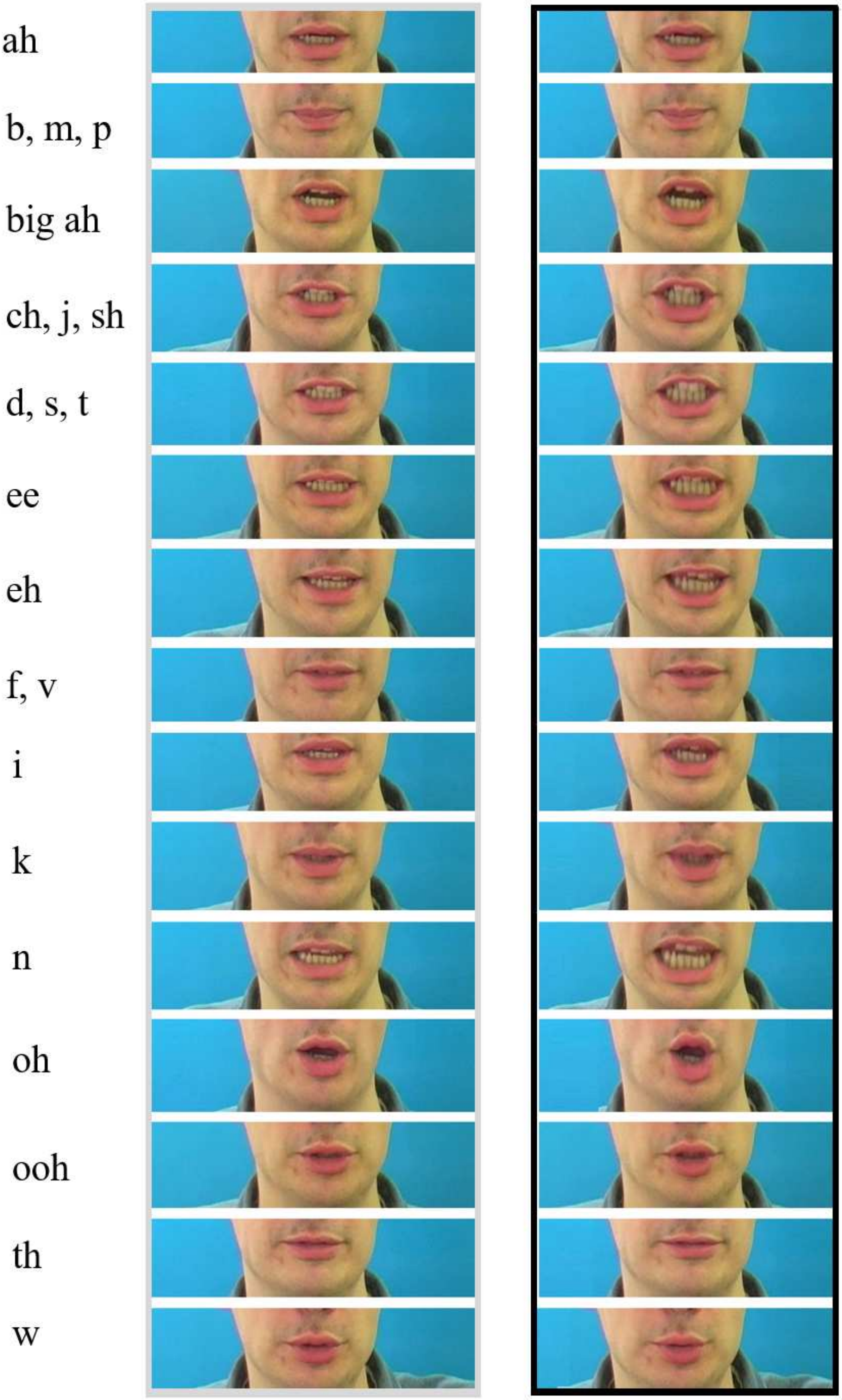

Figure 3: *Viseme classes extracted for the Grid data sets [53]. The first column represents the original viseme mouth shape while the second column represents the viseme mouth shape after applying the exaggeration effect ($\alpha = 2$). The British English Example Pronunciation dictionary was used for the phoneme notation.*

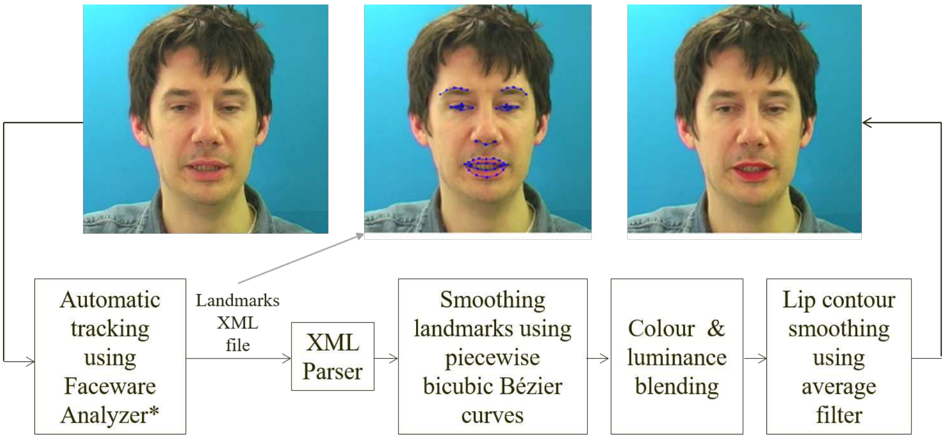
Automatic tracking using Faceware Analyzer*
Landmarks XML file
XML Parser
Smoothing landmarks using piecewise bicubic Bézier curves
Colour & luminance blending
Lip contour smoothing using average filter

Figure 4: *Lipstick-effect implementation stages [29].*

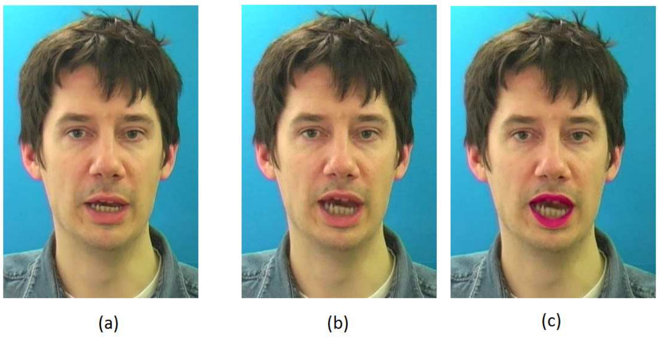

Figure 5: *(a) A frame in the V stimuli; the corresponding frames in (b) E and (c)* $E_{lipstick}$*;* $\alpha = 2$ *was used in the exaggeration effect applied to E and* $E_{lipstick}$.

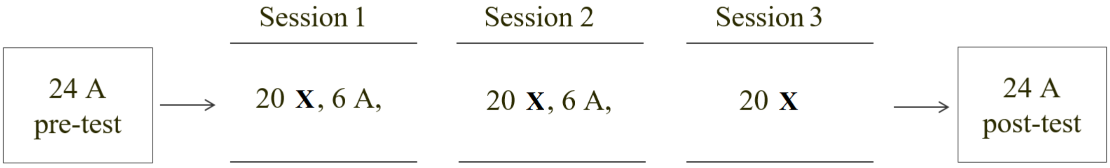

Figure 6: *The training method consisted of an audio-only pre-test followed by three training sessions where X = A, V, E, or $E_{lipstick}$, then an audio-only post-test session for determining the training gains.*

2.4. Procedure

Figure 6 illustrates the training methodology that was used in this study. In order to set a baseline level for each sub-group, all subjects took an audio-only pre-test of 24 A stimuli (12 stimuli repeated twice), presented in a random order. The assignment of a subject S to a training sub-group was done automatically when the subject finished the pre-test so as to establish a similar baseline across all sub-groups. This was done as follows. Assume that the subject's pre-test score is $S_{pre-test}$. The training software finds a sub-group X (X = A, V, E, or $\mathrm{E}_{lipstick}$) such that adding $S_{pre-test}$ to the set of X pre-test scores minimizes, makes no change, or makes the minimum increase to the standard deviation between the means of all sub-groups' pre-test scores. Subjects then underwent three training sessions. Each session consisted of an X (X = A, V, E, or $\mathrm{E}_{lipstick}$) training block that used ten X stimuli (repeated twice, then presented in a random order to give 20 X) and a test session of six A stimuli, except for session 3 that preceded the post-test. The two sets of six A presented in sessions 1 and 2 were used to track learning milestones for all sub-groups. After completing all training sessions, all subjects took an audio-only post-test using 24 A stimuli in order to assess their training gains.

Because the Grid corpus was used to provide stimuli, each subject's training/testing task was to identify the colour, letter, and digit that corresponded to the stimulus that was played; they then had to enter in these items using three presses of buttons on a labelled keyboard. During training (i.e. 20 X in session 1 ,2 and 3), after the subjects submitted their input for a given stimulus, that stimulus was then replayed with added subtitles to show the correct words, whether or not the input was correct. No such feedback was provided during testing (i.e. during the 24 A pre-test, the 6 A in session 1 and 2, and the 24 A post-test).

3. Results

Figure 7 summarises the main results of this experiment. Figure 7a shows the impact of speech modality (training stimuli) presented during the training on the recognition scores by the subject groups. Between groups, one-way ANOVA testing between the groups showed a significant difference between the V and E groups during the second training session ($F(3, 67) = 3.38$, $p = 0.02$). No significant difference was found between other groups in all training sessions. Within groups, repeated-measure ANOVA showed a significant difference between sentence-recognition scores in the E training sessions ($F(2, 40) = 9.987$, $p = 0.000$). A post-hoc pairwise comparison found a difference between sessions 1 and 3 ($p = 0.012$) and sessions 2 and 3 ($p = 0.000$). A significant difference was also found between the

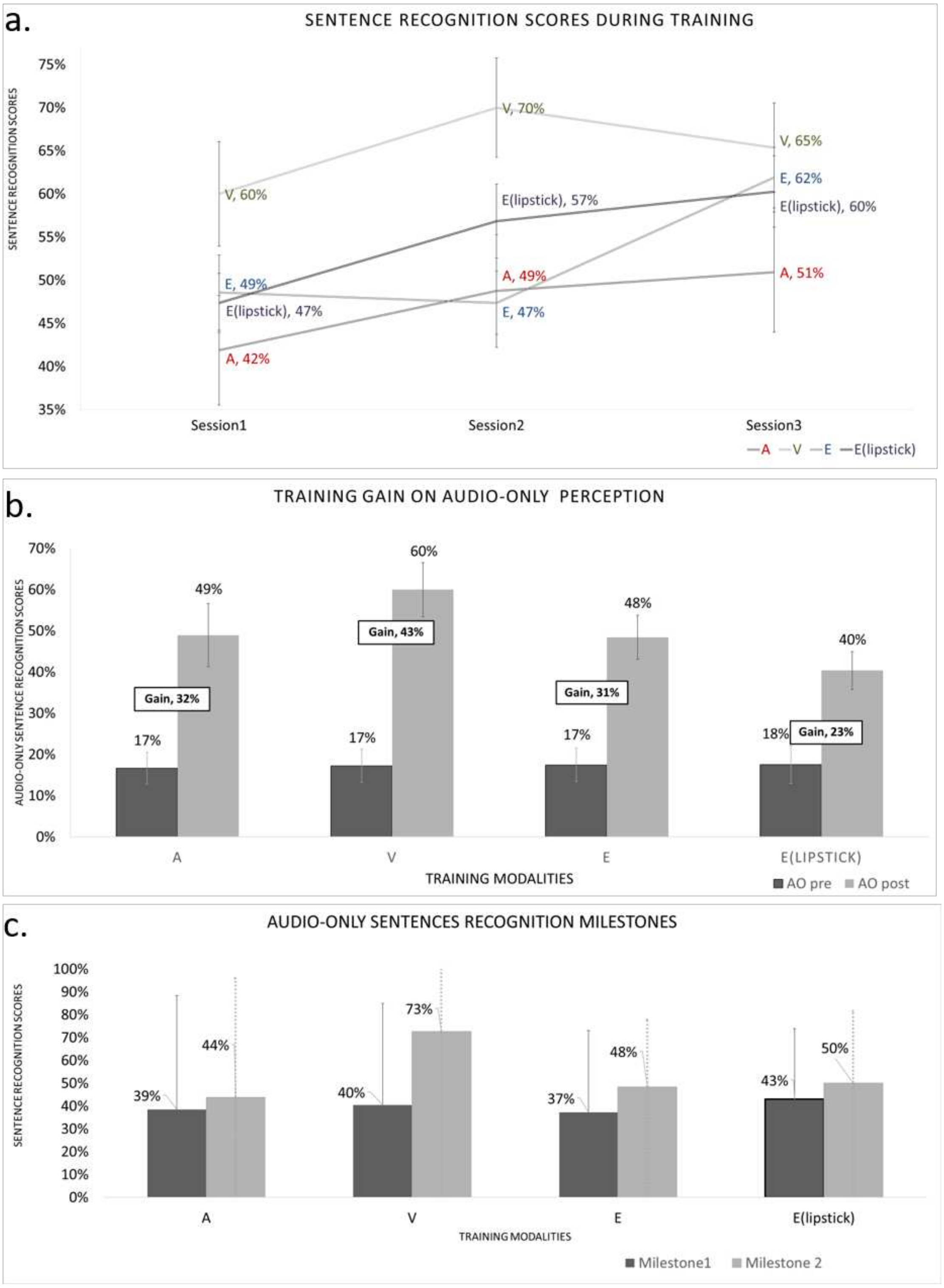

Figure 7: *Results for the A, V, E, and $E_{lipstick}$ subjects: (a) sentence recognition during training; (b) audio-only pre- and post-test mean identification scores and training gains (post-test results and pre-test results); (c) training impact on audio-only sentence recognition (learning milestones). Error bars = standard error.*

sentence-recognition scores in the E$_{lipstick}$ training sessions ($F(2, 36) = 3.38$, $p = 0.02$); the post-hoc test demonstrated a difference between sessions 1 and 3 ($p = 0.038$). No significant difference was found between the sentence-recognition scores in all sessions of the A and V training. Subjects who underwent the E and E$_{lipstick}$ training described the modified form of speech as incongruent audiovisual signals. More energy was observed in the visual signals than in audio signals (i.e. the video cues were more salient than the audio cues). This situation made audiovisual signals unintelligible at the start of the training.

Figure 7b shows the mean sentence-recognition scores that the A, V, E, and E$_{lipstick}$ subjects attained in their audio-only pre- and post-training tests as well as their mean training gains in auditory recognition. All sub-groups were formed with comparable pre-test scores (see Section 2.4). The V subjects achieved the highest post-test sentence recognition and training gains, although a one-way ANOVA test showed no significant difference between their sentence-recognition scores in post-testing ($F(3, 67) = 1.6$, $p = .19$) among the A, V, E, and E$_{lipstick}$ subjects. A Levene's test indicated unequal variances in training-gain scores ($F = 2.6$, $p = .041$), while Welch ANOVA testing reported no significant difference in training gains among all groups ($F(3, 35) = 0.00$, $p = 1.00$). Figure 7c shows the scores for the stimuli used in the 6 A of sessions 1 and 2 for all subgroups. These scores were used as Learning Milestones to track the audio-only skills for all sub-groups throughout the training. As Figure 7c shows, no significant differences were found in the session 1 post-testing among all groups (Milestone 1; $F(3, 67) = 0.14$, $p = .93$), while a significant difference was found between the A and V groups in session 2 post-testing (Milestone 2; $F(3, 67) = 2.91$, $p = 0.04$, post-hoc Tukey HSD test $p < .05$). Only the V subjects showed significant improvements in their recognition scores between Milestones 1 and 2 ($p = .002$). This may suggest that using unmodified visual signals in the training helped to speed up the V subjects' learning curve for auditory-only skills.

Confusion matrices (Figures 8 and 9) were also produced in order to understand the possible sources of confusion the subjects experienced in letter keywords recognition during the training and audio-only post-testing. Letter recognition was found to be the most challenging task for all subjects due to the need to select from a larger set with high variance (25 letters), and letters having shorter duration in terms of phonemes, and thus less information, as opposed to colours (4) and digits (10). During the training, the introduction of either modified or original visual signals improved the subjects' recognition of letters that contained vowel sounds (as well as bilabial, labiodental, and velar consonant sounds) compared to the audio-only training regime. Modified visual signals (E and E$_{lipstick}$ stimuli), however, improved the vowel recognition by 10 percent and the recognition of letters containing alveolar consonants by 7 percent. The A subjects' high confusion rates were observed in the pairs (A /**e**ɪ/, E /**iː**/), and in (N /e**n**/,M /e**m**/) and (T /**t**iː/, D /**d**iː/- G /**d**ʒiː/). These high rates may have indicated low abilities in processing vowels, nasality, and voicing cues that were distorted by the vocoder. High confusion rates were also observed among the V subjects at pairs (G /dʒ**iː**/, J /dʒ**eɪ**/) and (I /**a**iː/, E /iː/) which are visually similar letters. The E subjects' high confusion rates were observed at the pairs (A /**eɪ**/, O /**oʊ**/) and (P /**p**iː/,B /**b**iː/); this was likely the result of over-exaggerated mouth shapes.

After the training and during the audio-only post-testing, the V subjects outperformed the other group's auditory recognition of vowel letters, since they recognised 54 percent of the vowels compared with the A, E, and $\mathrm{E}_{lipstick}$ subjects, which scored 30, 41, and 46 percent, respectively. No significant difference in consonant recognition was found among the groups. The A subjects showed confusion between the (G /**d**ʒiː/, T /**ti**ː/), (V /viː/, O /oʊ/), (O /**oʊ**/, A /**eɪ**/) and (P /**pi**ː/, B /**bi**ː/) pairs; the V subjects with (C /siː/, T /tiː/), (T /**ti**ː/, G /**d**ʒiː/) and (V /v**iː**/, E /**iː**/); and the E and $\mathrm{E}_{lipstick}$ subjects with the (P /**pi**ː/, B /**bi**ː/) and (Q /kjuː/,T /tiː/) pairs.

4. Discussion

In previous research [21], we examined the effectiveness of automatically enhancing the appearance of a talker's lips in maximising the benefit of visual speech on improving the intelligibility of spectrally distorted speech in audiovisual training. In this paper, we went further by investigating the impact of exaggerating a talker's mouth kinematics in audiovisual speech. Because visual signals are a correlate of audio signals in audiovisual speech, exaggerating the visual signal alone in audiovisual speech will create incongruent inputs for listeners. Given this situation, the study reported in this paper investigated the subjects' ability to adapt to audiovisual mismatches after exaggerating visual speech. The study also investigated whether or not applying the exaggeration effect to audiovisual speech would improve the benefits of the visual signal.

Consistent with previous findings [18–21], we found that the introduction of unmodified audiovisual speech during auditory training improved the training gains in the auditory and audiovisual perception of spectrally distorted speech. Visual speech facilitation for speech-in-noise intelligibility [4] played a key role in improving the non-native subjects' audiovisual recognition rates for spectrally distorted sentences during the training. Using visual speech in training improved the subjects' auditory adaptation processes to spectrally distorted speech; the subjects were found to have significantly improved between learning milestones. This situation could reflect the impact of effective rapid perceptual learning.

Audiovisual conflict after-effect. After exposure to the audiovisual speech with exaggeration effect (E and $E_{lipstick}$), we found evidence of an audiovisual conflict after-effect. The subjects were sensitive to the conflict between the articulation energy and the vocal effort in the modified videos during the early training stages. They also underwent a recalibration process during audiovisual speech integration in order to overcome this conflict. This situation was supported by the adaptation profile of the modified audiovisual speech groups (Figure 7a), which reflected a dramatic increase in the audiovisual recognition rate during session 3. The increase reached a comparable level to that of the group that received congruent audiovisual speech signals (the V group), which indicates that the conflict impact became negligible to the E and $E_{lipstick}$ subjects after exposure. There is a difference, however, observed in the pace of the adaptation process between the E and $E_{lipstick}$ sub-groups; the $E_{lipstick}$ sub-group seemed to adapt faster as reflected by the increase in the recognition scores between sessions 1 and 2 in Figure 7a. This suggests that the lipstick filter may have an impact on accelerating the adaption process in the $E_{lipstick}$ subjects.

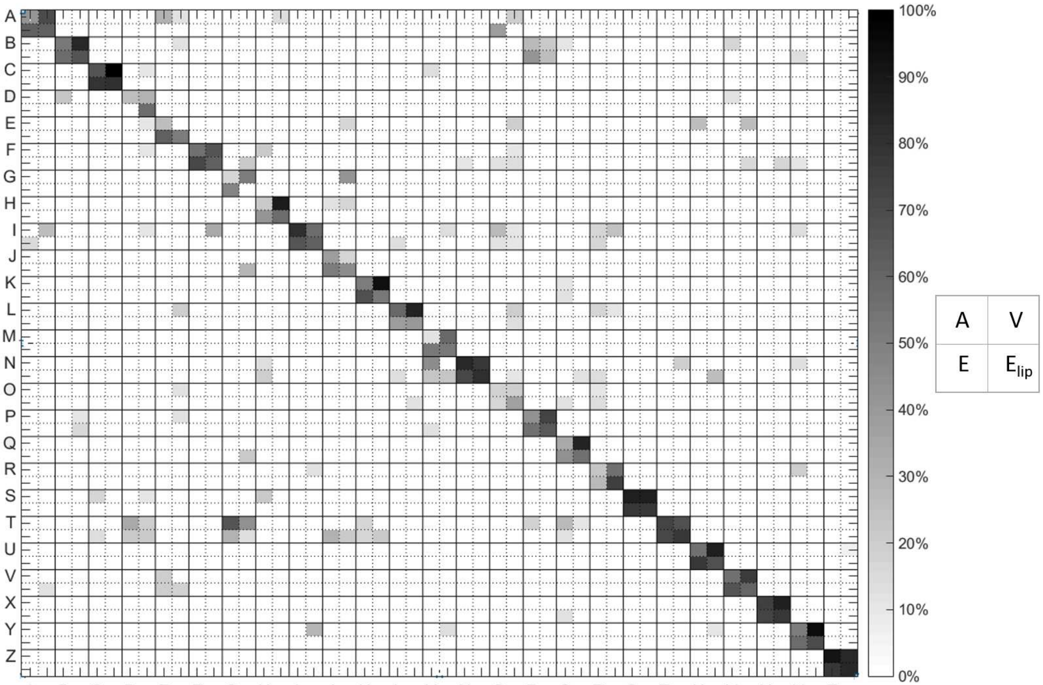

Figure 8: *Confusion matrix of letter-recognition scores during training for A, V, E, and $E_{lipstick}$ subjects.*

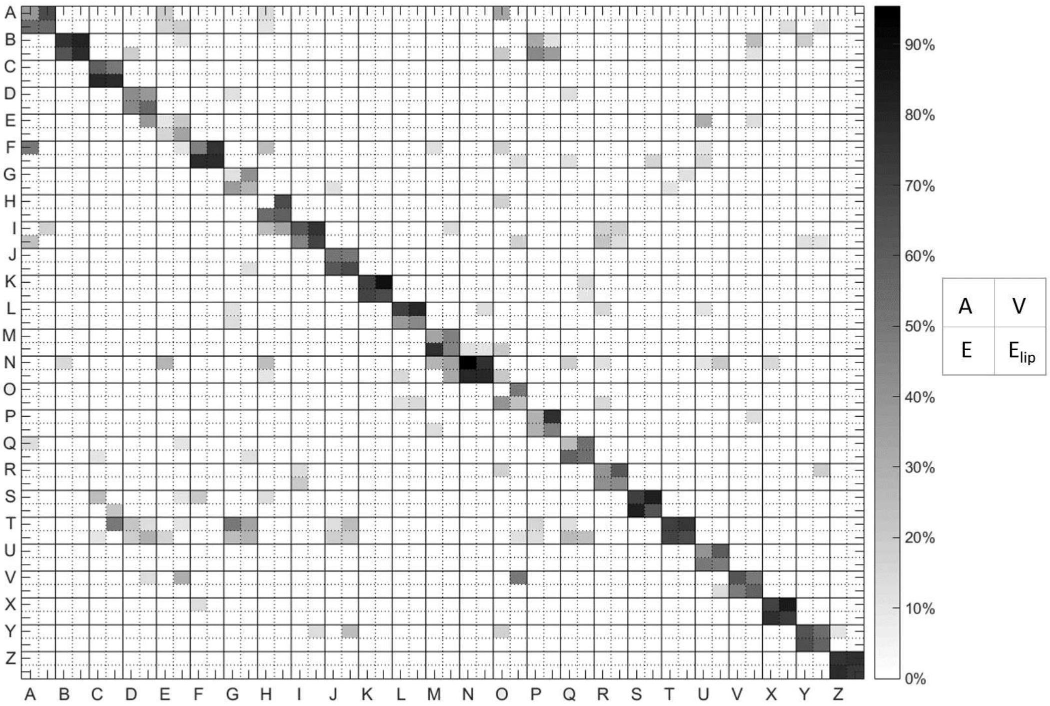

Figure 9: *Confusion matrix of letter-recognition scores during audio-only post-testing for A, V, E and $E_{lipstick}$ subjects.*

Impact of the exaggeration on audiovisual and auditory recognition. Exaggeration of the visual speech signal also improved the audiovisual recognition of vowels and alveolar consonants, which are included in 44 percent of the Grid letters. For the remainder of the Grid letters, the exaggeration of the visual signal showed a comparable benefit to the unmodified visual speech. However, it did not improve the subsequent auditory recognition. Those subjects who were trained with exaggerated speech attained training gains in auditory recognition that were comparable to the gains of those who had been trained with auditory-only speech. This situation indicates that the subjects did not make use of the modified visual signals to facilitate their auditory adaptation to the spectrally distorted speech. Our hypothesis is that the recalibration process the subjects underwent in order to adapt to the audiovisual conflict during the training was responsible for this. The recalibration process may have introduced additional cognitive load to the subjects, which in turn slowed down their auditory perceptual learning. It is thus difficult to judge whether or not modifying the audiovisual speech by exaggerating the visual signal can maximise the training gains in auditory recognition, since the subjects needed to undergo a recalibration process in order to adapt to the modified signals before they commenced the training.

Future work. Work is underway to improve the exaggeration effect in order to minimise audiovisual conflict. In the current method, visual signals are amplified by a constant value for all speech segments. This is unrealistic according to H&H theory [31], where energy fluctuates between hypo- and hyper-levels in articulation. To gain a better understanding of the articulatory behaviour of hyper-articulated speech, we conducted a small-scale experiment in which we compared our exaggeration method with real hyper-articulated speech, or Lombard speech [54]. Lombard speech is one of a suite of acoustic and phonetic modification techniques; these modifications include increases in fundamental frequency (F0) and speech levels, shifted first and second frequencies, increases in higher-frequency energy, and a flattened spectral tilt. At the phonetic level, Lombard speech is characterised by increased vowel duration as well as energy shifts among different classes of phonemes [55–68]. Visually, Lombard speech is characterised by greater global movement of the jaw as well as lip spreading, opening, closing, and protrusion [69, 70]. We chose Lombard speech as an example of hyper-articulation for this study, since it is a practical way to induce hyper-articulation in a controllable manner [55, 71]. We collected audiovisual recordings of four talkers who read sentences in normal conditions (i.e. with no background noise) and in Lombard conditions (in which they listened over headphones to babble speech presented at 80 dB SPL). The recordings that were made in normal mode were then modified to exaggerate the mouth movements of the talkers. The analysis of the articulatory measures (mouth width and height) at the phoneme level taken from all recordings (normal, Lombard, and exaggerated) revealed that hyper-articulated speech contained a number of features that were not present in our exaggeration method. The exaggeration method we used may simulate a similar effect to Lombard speech change in mouth height, but not mouth width. Future work will examine the effect of hyper-articulation and develop an exaggeration method derived from real Lombard speech data. Furthermore, as only female subjects were considered in this study, and because they are better lipreaders than male subjects [50], future work could also examine whether or not the exaggeration effect could help to facilitate lipreading in male subjects.

5. Conclusions

This paper has reported on a study that investigated the impact of introducing a modified visual signal (i.e. exaggerated mouth movements) to spectrally distorted audiovisual speech during auditory training. Our aim was to improve the audiovisual and subsequent audio-only speech recognition of spectrally distorted speech. We conducted tests on non-native listeners based on the reasoning that the hearing experience of such listeners for spectrally distorted speech is analogous to that of CI users. The results of the study suggest that the subjects who attained the ability to adapt to the mismatch between visual and audio signals did so as an after-effect of exposure to the exaggeration of the visual signal of audiovisual speech. As audiovisual conflict became negligible to subjects' after exposure, the results suggest the feasibility of applying enhancement effects on the visual signal alone in audiovisual speech, even if such enhancement may create incongruent audiovisual inputs.

This effect also improved the subjects' audiovisual recognition of certain phoneme classes. This situation also indicates the potential for this effect to be employed to facilitate face-to-face communication in a number of applications; for example, augmented reality solutions incorporated in communication or media platforms for listeners who are undergoing internal adverse conditions. The exaggeration effect, however, did not seem to show similar improvements in the subjects' subsequent auditory recognition. Their adaptation to the audiovisual conflict during the training may have played a role in impeding their facilitation of the visual signal in improving their subsequent auditory-only skills. Future work will involve devising improvements to the exaggeration method by seeking inspiration from real hyper-articulated speech data in order to minimise the audiovisual conflict.

Acknowledgements

This research has been supported by the Saudi Ministry of Education and King Saud University(KSU). We would like to thank Dr. Hessa Alsalamah, the Vice Dean of the College of Computer and Information Sciences (CCIS) in KSU and Dr. Heyam Albaity, the Head of the IT department in CCIS for facilitating the visit to university premises to conduct the experiment. We also thank Dr. Abir BenAbid and Dr. Areej Alwabil, members of Software and Knowledge Engineering Research Group (SKERG) in CCIS for offering the HCI lab as a place to conduct the experiment. We also thank students and staff from KSU who participated in the experiment. Many thanks to Prof. Bernd Möbius and the two anonymous reviewers for their helpful comments.

References

[1] D. W. Massaro, M. M. Cohen, Phonological context in speech perception, Perception & Psychophysics 34 (4) (1983) 338–348.

[2] L. D. Rosenblum, H. M. Saldaña, An audiovisual test of kinematic primitives for visual speech perception., Journal of Experimental Psychology: Human Perception and Performance 22 (2) (1996) 318.

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

[4] W. H. Sumby, I. Pollack, Visual contribution to speech intelligibility in noise, The Journal of the Acoustical Society of America 26 (2) (1954) 212–215.

[5] A. MacLeod, Q. Summerfield, Quantifying the contribution of vision to speech perception in noise, British journal of audiology 21 (2) (1987) 131–141.

[6] N. P. Erber, Interaction of audition and vision in the recognition of oral speech stimuli, Journal of Speech, Language, and Hearing Research 12 (2) (1969) 423–425.

[7] M. Middelweerd, R. Plomp, The effect of speechreading on the speech-reception threshold of sentences in noise, The Journal of the Acoustical Society of America 82 (6) (1987) 2145–2147.

[8] Q. Summerfield, A. MacLeod, M. McGrath, M. Brooke, Lips, teeth, and the benefits of lipreading, Handbook of Research on Face Processing (1989) 223–233.

[9] M. McGrath, An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces., Ph.D. thesis, University of Nottingham (1985).

[10] L. D. Rosenblum, J. A. Johnson, H. M. Saldana, Point-light facial displays enhance comprehension of speech in noise, Journal of Speech, Language, and Hearing Research 39 (6) (1996) 1159–1170.

[11] J. Kim, C. Davis, How visual timing and form information affect speech and non-speech processing, Brain and language 137 (2014) 86–90.

[12] N. P. Erber, Auditory training, Alex Graham Bell Assn for Deaf, 1982.

[13] A. R. Kaiser, K. I. Kirk, L. Lachs, D. B. Pisoni, Talker and lexical effects on audiovisual word recognition by adults with cochlear implants, Journal of Speech, Language, and Hearing Research 46 (2) (2003) 390–404.

[14] S. Desai, G. Stickney, F.-G. Zeng, Auditory-visual speech perception in normal-hearing and cochlear-implant listenersa), The Journal of the Acoustical Society of America 123 (1) (2008) 428–440.

[15] X. Li, K. Nie, N. S. Imennov, J. T. Rubinstein, L. E. Atlas, Improved perception of music with a harmonic based algorithm for cochlear implants, IEEE Transactions on Neural Systems and Rehabilitation Engineering 21 (4) (2013) 684–694.

[16] K. Nie, A. Barco, F.-G. Zeng, Spectral and temporal cues in cochlear implant speech perception, Ear and Hearing 27 (2) (2006) 208–217.

[17] D. S. Lazard, A.-L. Giraud, P. Barone, Multisensory interactions in auditory cortex and auditory rehabilitation in deafness (2013) 217–236.

[18] L. E. Bernstein, E. T. Auer Jr, J. Jiang, S. P. Eberhardt, Auditory perceptual learning for speech perception can be enhanced by audiovisual training, Frontiers in Neuroscience 7 (2013) 34.

[19] M. Pilling, S. Thomas, Audiovisual cues and perceptual learning of spectrally distorted speech, Language and speech (2011).

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

[21] N. Alghamdi, S. Maddock, G. J. Brown, J. Barker, A comparison of audiovisual and auditory-only training on the perception of spectrally-distorted speech, in: ICPhS: The International Congress of Phonetic Scences, 2015.

[22] K. Lander, C. Capek, Investigating the impact of lip visibility and talking style on speechreading performance, Speech Communication 55 (5) (2013) 600–605.

[23] R. Campbell, T.-J. Mohammed, Speechreading for information gathering: a survey of scientific sources, deafness cognition and language (dcal) research centre, division of psychology and language sciences, university college london, 2010.

[24] V. Scott, Belonging, Flying Fingers Club Series, Kendall Green Publications, Gallaudet University Press, 1987.

[25] J. E. Preminger, H.-B. Lin, M. Payen, H. Levitt, Selective visual masking in speechreading, Journal of Speech, Language, and Hearing Research 41 (3) (1998) 564–575.

[26] Y. Kitano, B. M. Siegenthaler, R. G. Stoker, Facial hair as a factor in speechreading performance, Journal of Communication Disorders 18 (5) (1985) 373–381.

[27] H. Kaplan, S. Bally, C. Garretson, Speechreading: A Way to Improve Understanding, Gallaudet University Press, 1985.

[28] N. Daly, J. Bench, H. Chappell, Gender differences in visual speech variables, Journal-Academy of Rehabilitative Audiology 30 (1997) 63–76.

[29] N. Alghamdi, S. Maddock, G. J. Brown, J. Barker, Investigating the impact of artificial enhancement of lip visibility on the intelligibility of spectrally-distorted speech., in: FAAVSP : The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing, Vienna, 2015.

[30] P. Edwards, C. Landreth, E. Fiume, K. Singh, Jali: An animator-centric viseme model for expressive lip synchronization, ACM Trans. Graph. 35 (4) (2016) 127:1–127:11.

[31] B. Lindblom, Explaining phonetic variation: A sketch of the h&h theory, Speech Production and Speech Modelling 55 (2012) 403.

[32] B. Theobald, R. Harvey, S. Cox, G. Owen, C. Lewis, Lip-reading enhancement for law enforcement, in: SPIE conference on Optics and Photonics for Counterterrorism and Crime Fighting, 2006.

[33] P. Bertelson, J. Vroomen, B. De Gelder, Visual recalibration of auditory speech identification a mcgurk aftereffect, Psychological Science 14 (6) (2003) 592–597.

[34] R. I. Bermant, R. B. Welch, Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition., Perceptual and Motor Skills.

[35] H. M. Saldaña, L. D. Rosenblum, Visual influences on auditory pluck and bow judgments, Perception & Psychophysics 54 (3) (1993) 406–416.

[36] B. De Gelder, J. Vroomen, The perception of emotions by ear and by eye, Cognition & Emotion 14 (3) (2000) 289–311.

[37] J. Vroomen, M. Keetels, B. De Gelder, P. Bertelson, Recalibration of temporal order perception by exposure to audio-visual asynchrony, Cognitive Brain Research 22 (1) (2004) 32–35.

[38] W. Fujisaki, S. Shimojo, M. Kashino, S. Nishida, Recalibration of audiovisual simultaneity, Nature Neuroscience 7 (7) (2004) 773–778.

[39] B. Chen, F. Dachille, A. Kaufman, Forward image mapping (1999) 89–96.

[40] T. Leyvand, D. Cohen-Or, G. Dror, D. Lischinski, Data-driven enhancement of facial attractiveness, ACM Transactions on Graphics (TOG) 27 (3) (2008) 38.

[41] S. Melacci, L. Sarti, M. Maggini, M. Gori, A template-based approach to automatic face enhancement, Pattern Analysis and Applications 13 (3) (2010) 289–300.

[42] V. Kitanovski, E. Izquierdo, Augmented reality mirror for virtual facial alterations, 18th IEEE International Conference on Image Processing (2011) 1093–1096.

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

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

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

[46] P. Assmann, Q. Summerfield, The perception of speech under adverse conditions (2004) 231–308.

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

[48] V. Hazan, A. Sennema, A. Faulkner, Audiovisual perception in l2 learners., in: Interspeech, 2002.

[49] D. M. Hardison, Acquisition of second-language speech: Effects of visual cues, context, and talker

variability, Applied Psycholinguistics 24 (04) (2003) 495–522.

[50] J. Dancer, M. Krain, C. Thompson, P. Davis, et al., A cross-sectional investigation of speechreading in adults: effects of age, gender, practice, and education., The Volta Review.

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

[52] T. Cootes, An introduction to active shape models, Image Processing and Analysis (2000) 223–248.

[53] R. Alghady, Y. Gotoh, S. Maddock, Analysis of visemes in the grid corpus, in: UKSpeech 2016 Conference at the University of Sheffield.

[54] E. Lombard, Le signe de l'elevation de la voix, Ann. Maladies Oreille, Larynx, Nez, Pharynx 37 (101-119) (1911) 25.

[55] J. Simko, S. Benu, M. Vainio, Hyperarticulation in lombard speech: A preliminary study, in: Proceedings of the 7th International Conference on Speech Prosody, 2014.

[56] J.-C. Junqua, The lombard reflex and its role on human listeners and automatic speech recognizers, The Journal of the Acoustical Society of America 93 (1) (1993) 510–524.

[57] J.-C. Junqua, S. Fincke, K. Field, The lombard effect: A reflex to better communicate with others in noise 4 (1999) 2083–2086.

[58] Y. Lu, M. Cooke, Speech production modifications produced by competing talkers, babble, and stationary noise, The Journal of the Acoustical Society of America 124 (5) (2008) 3261–3275.

[59] Y. Lu, Production and perceptual analysis of speech produced in noise, Ph.D. thesis, University of Sheffield (2010).

[60] D. K. Amazi, S. R. Garber, The lombard sign as a function of age and task, Journal of Speech, Language, and Hearing Research 25 (4) (1982) 581–585.

[61] C. Davis, J. Kim, K. Grauwinkel, H. Mixdorff, Lombard speech: Auditory (a), visual (v) and av effects (2006) 248–252.

[62] M. Cooke, S. King, M. Garnier, V. Aubanel, The listening talker: A review of human and algorithmic context-induced modifications of speech, Computer Speech & Language 28 (2) (2014) 543–571.

[63] J. Kim, C. Davis, Comparing the consistency and distinctiveness of speech produced in quiet and in noise, Computer Speech & Language 28 (2) (2014) 598–606.

[64] J. Robert-Ribes, J.-L. Schwartz, T. Lallouache, P. Escudier, Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of french oral vowels in noise, The Journal of the Acoustical Society of America 103 (6) (1998) 3677–3689.

[65] C. Davis, A. Sironic, J. Kim, Perceptual processing of audiovisual lombard speech.

[66] C. Davis, J. Kim, Is speech produced in noise more distinct and/or consistent?, Speech Science and Technology (2012) 46–49.

[67] J. Kim, A. Sironic, C. Davis, Hearing speech in noise: Seeing a loud talker is better, Perception-London 40 (7) (2011) 853.

[68] M. D. Skowronski, J. G. Harris, Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments, Speech Communication 48 (5) (2006) 549–558.

[69] M. Garnier, L. Bailly, M. Dohen, P. Welby, H. Loevenbruck, An acoustic and articulatory study of lombard speech: Global effects on the utterance, in: The Ninth International Conference on Spoken Language Processing (Interspeech/ICSLP), 2006.

[70] J. Kim, C. Davis, G. Vignali, H. Hill, A visual concomitant of the lombard reflex. (2005) 17–22.

[71] J. Šimko, Š. Beňuš, M. Vainio, Hyperarticulation in lombard speech: Global coordination of the jaw, lips and the tongue, The Journal of the Acoustical Society of America 139 (1) (2016) 151–162.