Reducing Viseme Confusion In Speech-Reading

Benjamin M. Gorman
Computing, School of Science and Engineering
University of Dundee, Dundee, Scotland
b.gorman@dundee.ac.uk

Abstract

Speech-reading is an invaluable technique for people with hearing loss or those in adverse listening conditions (e.g., in a noisy restaurant, near children playing loudly). However, speech-reading is often difficult because identical mouth shapes (visemes) can produce several speech sounds (phonemes); there is a one-to-many mapping from visemes to phonemes. This decreases comprehension, causing confusion and frustration during conversation. My doctoral research aims to design and evaluate a visualisation technique that displays textual representations of a speaker's phonemes to a speech-reader. By combining my visualisation with their pre-existing speech-reading ability, speech-readers should be able to disambiguate confusing viseme-to-phoneme mappings without shifting their focus from the speaker's face. This will result in an improved level of comprehension, supporting natural conversation.

Problem and Motivation

Speech-reading (often called lip-reading) refers to using visual information about the movements of a speaker's lips, teeth, and tongue to understand what they are saying. Speech-reading is commonly used by those who are deaf and those with hearing loss to understand speech (Campbell, Dodd, & Burnham, 1998), and it has been shown that those with typical hearing also speech-read (albeit subconsciously) to help them understand others (Goldstein, 2013).

An atomic unit of a language's phonology is called a phoneme. Phonemes are combined to form words. For example, **/b/**, **/æ/**, and **/t/** are the phonemes for the word "bat". There are 48 commonly-recognised phonemes in English (Shoup, 1980). For each phoneme, a speaker's lips, teeth, and tongue produce a visual representation known as a viseme. Speech-readers reverse this mapping (i.e., viseme-to-phoneme) to determine the sounds a speaker is making, which helps them understand the words being said. For example, in English, **/l/** and **/r/** are acoustically similar (especially in clusters, such as "grass" vs. "glass"), but visually distinct, so speech-readers use this difference to determine if a speaker has said **/l/** or **/r/**.

However, the viseme-to-phoneme mapping is often a 'one-to-many' relationship, in which a single viseme maps to a number of phonemes (Lucey, Martin, & Sridharan, Confusability of phonemes grouped according to their viseme classes in noisy environments, 2004). For example, the phoneme **/v/**, is a voiced sound distinct from **/f/**, which is not voiced. However, the viseme for **/v/** is almost identical to the viseme for **/f/**, making the words "fan" and "van" difficult to distinguish for a speech-reader. As a result, speech-reading alone is often not sufficient to fully understand what a speaker is saying, and this can result in confusion, frustration, and reduced conversational confidence for speech-readers (Campbell, Dodd, & Burnham, 1998). This can lead to individuals feeling isolated or excluded during social activities, undermining their self-esteem, and becoming concerned about fulfilling their potential at work (Ringham).

Related Work

A number of techniques have been developed to overcome the challenges presented by viseme-to-phoneme confusion during speech-reading. Sign languages (e.g., American Sign Language, British Sign Language) use hand and arm gestures and body language to facilitate communication. Cued Speech (Cornett, 1967) combines a small number of hand shapes, known as cues (representing consonants) in different locations near the mouth (representing vowels), as a supplement to speech-reading. However, for both sign language and cued speech to be effective they require both parties in a conversation to be fluent in the technique used, limiting the situations where these techniques are applicable.

Speech can be visualised by showing the intensity of sound at different frequencies over time. This can be shown graphically in a spectrogram, where time is on the X axis, frequency is on the Y axis, and intensity maps to colour. Spectrograms are used by linguists to identify words phonetically, although becoming competent can take considerable training (Greene, Pisoni, & Carrell, 1984). Vocsyl (Hailpern, Karahalios, DeThorne, & Halle, 2010) is a software system that provides visual feedback of speech. Pietrowicz and Karahalios built upon this work by adding colour mappings to the visualisation to represent phonological detail (Pietrowicz & Karahalios, 2013). Vocsyl and its extension were not designed to supplement speech-reading, therefore Vocsyl can lead to multiple words having similar visual representations; words within the same viseme groups such as "fan" and "van" are coded with the same colours because **/f/** and **/v/** both have the same phoneme class (fricative).

Closed captioning (captioning, subtitling) displays the audio of a television programme as text on the TV screen, providing access to the speech and sound effects to individuals who are deaf or hard-of-hearing. Caption creation typically relies to some extent on human transcription. Captions also require the viewer to split their attention between reading and watching the video content (or the speaker's face); one eye-tracking study found that participants spent around 84% of their viewing time focussed exclusively on captions (Jensema, Danturthi, & Burch, 2000).

There have also been attempts at training speech-readers by showing computer-generated facial models, supplemented with additional cues. Lip Assistant (Xie, Wang, & Liu, 2006) is a system which uses a video synthesizer to generate magnified video-realistic mouth animations of the speaker's lips. The rendered mouth animations are superimposed to the bottom left corner of the original video. iBaldi (Massaro, Cohen, Schwartz, Vanderhyden, & Meyer, 2013), is an iOS application that shows a computer animated face and transforms speech into visual cues to supplement speech-reading. The cues are three coloured discs, showing nasality (red), friction (white), and voicing (blue), which appear when a phoneme from a corresponding group is presented. The cues are located near the computer generated face's mouth.

These visualisation techniques have limitations that substantially restrict their value to speech-readers: **1) Low Accuracy** – many of these techniques have accuracy rates similar to unassisted speech-reading. In preliminary studies I have found that in a constrained word recognition task with accurate transcript and timing, many of these techniques do not allow significant benefits over visual only speech-reading. **2) High Training** – many of these techniques require substantial amounts of training, e.g., iBaldi evaluations used 30 to 50 hours of training (Massaro, Cohen, Schwartz, Vanderhyden, & Meyer, 2013). **3) Split-Attention** – many of these techniques require split-attention between the technique and the speaker which can negatively impact on the flow of natural conversation.

Proposed Solution

To address the limitations of current techniques, I am developing PhonemeViz, a visualisation technique, that displays a subset of a speaker's phonemes to the speech-reader. The visualisation aims to reduce viseme confusion which occurs at the start of words. PhonemeViz places the most recently spoken initial phoneme on a circular opaque background and is positioned at the side of a speaker's lips to allow the speech-reader to focus on the speaker's eyes and lip movements while monitoring changes in PhonemeViz's state using their peripheral vision. Through a combination of looking at the visualisation and their ability to speech-read, speech-readers should be able to attend to the speaker's face while being able to disambiguate confusing viseme-to-phoneme mappings, therefore improving understanding during conversation.

The end goal is to display the visualisation on a transparent head mounted display (as shown in Figure 10), such as the Epson Moverio glasses or the Microsoft Hololens, as a visual augmentation of speech. The visualisation could also be superimposed onto video content in a similar manner to subtitles as a form of training material for learning speech-reading or as a way to access media.

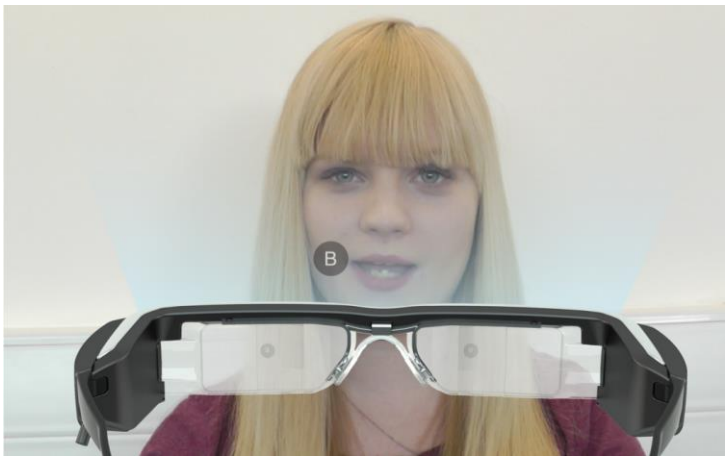

Figure 10: A rendering of the current iteration of PhonemeViz, showing what would be displayed on the Epson Moverio Glasses for the word "Bat".

Progress and Future Work

My PhD research is comprised of three main research stages. The first stage, which was completed in Spring 2014, was a preliminary quantitative study that compared an initial PhonemeViz prototype to other related techniques. The second, and current, stage is comprised of gathering data to inform future development. The third stage is the final evaluation of the visualisation technique.

Stage 1: Preliminary Study

Prototype Design

The initial prototype visualisation of PhonemeViz focused on reducing viseme confusion, when it occurs at the start of words. This version placed consonant phonemes in a semi-circular arrangement, with an arrow beginning from the centre of this semi-circle pointing at the last heard consonant phoneme to provide persistence. PhonemeViz was designed to be positioned at the side of a speaker's face, beginning at the forehead and ending at the chin. This should allow

the speech-reader to focus on the speaker's eyes and lip movements while monitoring changes in PhonemeViz's arrow using their peripheral vision.

Beginning with 48 phonemes, the vowel phonemes were removed (as vowels are more easy to distinguish visually) which left 29 phonemes. This was further reduced to a set of 22 phonemes by simplifying similar phonemes into one representation. We used the viseme categories to identify locations within the semi-circle that were spatially distributed for each entry within a viseme category. To facilitate learning, and hence require less focus on the visualisation, we alphabetically-ordered each phoneme representation within a viseme from the top of the semi-circle to the bottom.

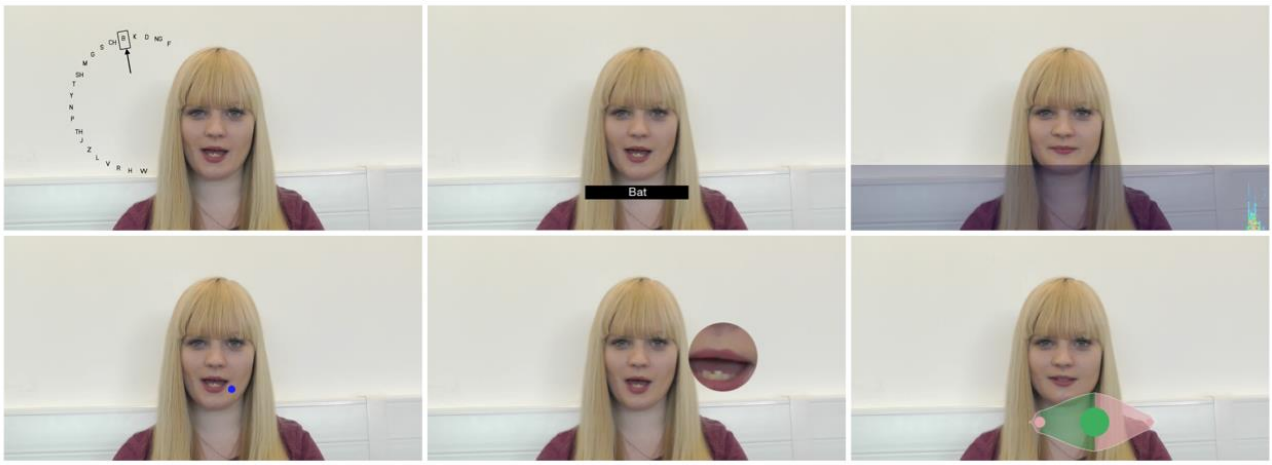

Figure 11: Each visualization technique composited over video of the word `Bat'; Top Row: PhonemeViz (left), Captions (middle) and Spectrogram (right), Bottom Row: iBaldi (left), LipMag (middle) and VocSyl (right).

Evaluation

The initial prototype was evaluated in-lab against five other visualisation techniques (Captions, iBaldi (Massaro, Cohen, Schwartz, Vanderhyden, & Meyer, 2013), Lip-Magnification (Xie, Wang, & Liu, 2006), Spectrogram and VocSyl (Hailpern, Karahalios, DeThorne, & Halle, 2010) see Figure 11), as well as a no visualisation control (None). All participants had typical hearing, however all but one reported that they used speech-reading in day-to-day life. The evaluation used a repeated measures design and participants attended two study sessions over consecutive days.

Table 2: Experimental words with their corresponding viseme group. Target words identified with (*).

Viseme Group	Word 1	Word 2	Word 3
/p/	Pat	Mat	**Bat***
/t/	Sun	**Done***	Tonne
/k/	Kill	**Gill***	Nil
/ch/	Chill	Shill	**Jill***
/p/	Banned	Manned	**Panned***
/k/	Light	Night	**Kite***
/t/	**Zone***	Tone	Sewn

Task & Stimuli

The words for the evaluation were chosen by looking at the phoneme to viseme table in (Lucey, Martin, & Sridharan, Proc. of Australian Int. Conf. on Speech Science & Tech, 2004). Three words

were chosen for each viseme group that were similar, apart from the initial consonant phoneme. In total the evaluation uses four groups **/p/,/t/,/k/,/ch/**, with three (**/p/**, **/k/** and **/t/**) being repeated, albeit with different words as shown in Table 2.

The speaker was recorded saying each word three times, to help capture realistic subtle variations in speech, as would be the case in day-to-day conversation and reduced participants' familiarity with each video. Each technique was implemented as an openFrameworks (openframeworks.cc) application and used the ofxTimeline (github.com/YCAMInterlab/ofxTimeline) add-on. This add-on allowed us to add time-based triggers to a video. Using a list of phonemes and their features, we loaded each word video into an application for each technique, overlaying the technique's visualisation onto the image sequence and exported the resulting video using an additional openFrameworks add-on ofxVideoRecorder (github.com/timscaffidi/ofxVideoRecorder).

Stimuli were presented without audio to control for hearing ability as used in previous studies [14,15]. Participants used each technique to determine when a particular word in a group of words from the same viseme group was spoken. The participant was told to press the spacebar when they thought the speaker had said the target word. The order of each technique and word group were counterbalanced.

Results

We calculated F_1 Scores for each condition per participant using their precision and recall values calculated from their responses. The F_1 score was our dependent variable, while our independent variables were session and technique used. Since there was no significant effect found across sessions, we chose to regard the first session as training and focus our analysis on data from the second session. Our results are shown in Table 2.

PhonemeViz enabled all participants to achieve perfect F_1 scores (100% accuracy, 0% errors), which was significantly higher than all other techniques except Captions. None of the remaining techniques performed significantly differently than having no technique at all, indicating that they possibly did not offer much assistance in our study's speech-reading task. When using PhonemeViz, participants reported lower Mental Demand, Temporal Demand, Effort, and Frustration, as well as higher Performance, than all other techniques except Captions. Participants rated PhonemeViz as their second-most preferred option when asked if they would use the technique again (Captions were ranked first). There was a sizable gap in average rankings between PhonemeViz and the third-best technique (VocSyl).

Table 3: Mean F1 scores & Standard Error for each technique.

Technique	Mean F_1 Score	Standard Error
None	0.44	0.03
Spectrogram	0.36	0.07
VocSyl	0.55	0.09
Captions	0.99	0.01
LipMagnification	0.37	0.04
iBaldi	0.42	0.07
PhonemeViz	1.00	0

Limitations

The results of the comparative evaluation show that PhonemeViz allowed all participants to achieve 100%-word recognition (showing successful disambiguation), and participants rated the technique well in subjective and qualitative feedback. This demonstrates that visualising phonemes can improve visual only speech-reading in constrained word recognition tasks. However, there are four limitations with the preliminary study: **L1)** – Our participants were not individuals who relied on speech-reading for communication, therefore this preliminary study gives no insights into the performance of speech-readers. **L2)** – In this study, PhonemeViz only shows consonant phonemes. As part of expanding the set of phonemes, we will revisit how to distribute and highlight the phonemic character representations in the periphery, to ensure we do not lose the strengths demonstrated. **L3)** –The results do not indicate if visualising phonemes would result in better comprehension with sentences or during natural conversation as this depends on context. **L4)** – The results do not demonstrate if the visualisation detracted from the participants' ability to speech-read, as we do not know to what extent the participants were splitting their attention between looking at the face and the visualisation.

Stage 2: PhonemeViz Design

The participants of the preliminary study were not individuals who rely on speech-reading and now we know that the technique has potential, it is necessary to involve expert users in the design phase (Addresses **L1**). First, I will conduct one-to-one interviews with six or more of the 21 speech-reading tutors in Scotland (Armstrong, 2015), to learn more about the challenges of teaching and learning speech-reading. These interviews will be transcribed and thematically analysed to generate requirements that will be used to inform the design of PhonemeViz. Second, an online survey targeting speech-readers, will be conducted to gather their opinions on these requirements. This will lead to a design workshop with speech-readers and speech-reading tutors to gather their collective insights on the design of PhonemeViz (Addresses **L2**). Following the design workshop, an interactive, online study of several prototype visualisations will be conducted to gain quantitative data from a large number of my target user group (Addresses **L1,L3**). The results of this study will further refine the design of PhonemeViz.

Stage 3: Visualisation Evaluation

The final stage will be a full evaluation of the visualisation technique with speech-readers. I will conduct an in-lab study of the technique using sentence stimuli and participants from local speech-reading classes. The goal of the evaluation will be to see whether PhonemeViz allows speech-readers to have greater comprehension of sentences compared to having no assistance (Addresses **L1,L3,L4**).

Expected Contributions

My research will make three contributions to the fields of information visualisation and human computer interaction. First, I will introduce PhonemeViz, a new visualisation technique to support speech-readers. Second, I will conduct an evaluation of this visualisation with the target user group. Third, I will present a summary of design requirements for visual augmentations to support speech-reading based on qualitative findings from the target user group.

On a more general level, my research will provide answers to the question of whether speech-reading can be augmented successfully with additional information in a way that is not distracting or detrimental to natural conversation.

References

[1] R. Campbell, B. Dodd and D. K. Burnham, Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech, vol. 2, Psychology Press, 1998.

[2] E. Goldstein, Sensation and perception, Cengage Learning, 2013.

[3] J. E. Shoup, "Phonological aspects of speech recognition," *Trends in speech recognition,* pp. 125--138, 1980.

[4] P. Lucey, T. Martin and S. Sridharan, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in *Proc. of Australian Int. Conf. on Speech Science & Tech*, 2004.

[5] L. Ringham, "Not Just Lip Service," Action On Hearing Loss, [Online]. Available: http://www.actiononhearingloss.org.uk/notjustlipservice.aspx.

[6] R. O. Cornett, "Cued speech," *Am. Ann. Deaf.,* vol. 112, no. 1, pp. 3--13, 1967.

[7] B. G. Greene, D. B. Pisoni and T. D. Carrell, "Recognition of speech spectrograms," *JASA,* vol. 76, no. 1, pp. 32--43, 1984.

[8] J. Hailpern, K. Karahalios, L. DeThorne and J. Halle, "Vocsyl: Visualizing syllable production for children with ASD and speech delays," in *Proc. ASSETS '10*, 2010.

[9] M. Pietrowicz and K. Karahalios, "Sonic shapes: Visualizing vocal expression," in *ICAD 2013*, 2013.

[10] C. J. Jensema, R. S. Danturthi and R. Burch, "Time spent viewing captions on television programs," *Am. Ann. Deaf.,* vol. 145, no. 5, pp. 464--468, 2000.

[11] L. Xie, Y. Wang and Z.-Q. Liu, "Lip Assistant: Visualize Speech for Hearing Impaired People in Multimedia Services," in *Proc. SMC'06*, 2006.

[12] D. W. Massaro, M. M. Cohen, W. Schwartz, S. Vanderhyden and H. Meyer, "Facilitating Speech Understanding for Hearing-Challenged Perceivers in Face-to-Face Conversation and Spoken Presentations," *ICTHP,* 2013.

[13] P. Lucey, T. Martin and S. Sridharan, "Proc. of Australian Int. Conf. on Speech Science & Tech," in *Confusability of phonemes grouped according to their viseme classes in noisy environments*, 2004.

[14] L. E. Bernstein, P. E. Tucker and M. E. Demorest, "Speech perception without hearing," *Perception & Psychophysics,* vol. 62, no. 2, pp. 233--252, 2000.

[15] E. T. Auer and L. E. Bernstein, "Enhanced visual speech perception in individuals with early-onset hearing impairment," *J Speech Lang Hear Res,* vol. 50, no. 5, pp. 1157--1165, 2007.

[16] L. Armstrong, "On everybody's lips," Scottish Lipreading Strategy Group, 2015. [Online]. Available: http://www.scotlipreading.org.uk/files/1914/2686/1587/On_everybodys_lips_-_report.pdf.

[17] N. A. Altieri, D. B. Pisoni and J. T. Townsend, "Some normative data on lip-reading skills (L)," *JASA,* vol. 130, pp. 1-4, 2011.

[18] C. R. Berger and R. J. Calabrese, "Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication," in *HUM. COMMUN. RES.*.

[19] C. R. Lansing and G. W. McConkie, "Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences," *Percept. Psychophys.*, vol. 65, pp. 536--552, 2003.

About the Author:

Benjamin Gorman is a 2nd year PhD student at the University of Dundee in Dundee, Scotland and is supervised by Dr. David R. Flatla. He received a BSc in Applied Computing from the University of Dundee in 2012. In 2014, he started his PhD exploring how visualising aspects of speech, particularly phoneme information, can improve speech-reading. At Dundee he is a member of DAPRlab (Digitally Augmented Perception Research Lab). Visit his personal website at www.benjgorman.com.