ASSETS 2015 Doctoral Consortium Project Summary: Reducing Viseme Confusion In Speech-Reading

Benjamin M. Gorman
School of Computing, University of Dundee
Dundee, Scotland, DD1 4HN
b.gorman@dundee.ac.uk

ABSTRACT

Speech-reading is an invaluable technique for people with hearing impairments or those listening 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. In this doctoral consortium, I will present PhonemeViz as a solution to this problem. PhonemeViz is a phoneme visualisation technique that allows a speech-reader to disambiguate confusing viseme-to-phoneme mappings by visualising phoneme information using movement within their peripheral vision. Early results suggest that PhonemeViz can enable speech-readers to overcome the barriers they face when speaking with other people.

1. PROBLEM & 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 impairments to understand speech [2], and has been shown that those with typical hearing also speech-read (albeit subconsciously) to help them understand others [3]. When speech reading, each speech sound (phoneme) corresponds to a facial and mouth position (viseme). For example, acoustically speaking in English, /l/ and /r/ can be quite similar (especially in clusters, such as 'grass' vs. 'glass'), yet visual information can show a clear contrast.

An atomic unit of a language's phonology is called a *phoneme*, which is combined with other phonemes to form words. For example, /b/, /a/, and /t/ are the phonemes comprising the word 'bat'. There are 48 commonly-recognized phonemes in English [8]. For each phoneme, a speaker's lips, teeth, and tongue produce a visual representation of that phoneme called 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.

Unfortunately for speech-readers, the viseme-to-phoneme mapping is a 'one-to-many' relationship, in which a single viseme can map to numerous phonemes [6]. For example, acoustically /v/ is voiced, and /f/ is not voiced however, the viseme for /v/ is essentially identical to the viseme for /f/. As a result, speech-reading alone is often not sufficient to fully understand what a speaker is saying, and this commonly results in confusion, frustration, and reduced conversational confidence for speech-readers [2].

It has been shown that the amount of talking in the initial stages of an interpersonal relationship can have a significant effect on the strength of that relationship [1]; as the amount of verbal and nonverbal communication increase, the levels of uncertainty of both parties decreases, which leads to higher levels of intimacy and liking. However, barriers to communication may result in a stagnant career, social isolation and decrease in overall life satisfaction. How well and how willing we are to communicate, and the degree of our apprehension about the process of communicating have profound effects on our lives [2].

To help speech-readers, several visualisation techniques have been proposed, however these have limitations that substantially restrict their value to speech-readers. In spite of being specifically designed to improve speech-reading, most of these techniques have accuracy rates similar to unassisted speech-reading. Captions have been shown to have high accuracy, however these usually require the speech-reader to not attend to the speaker's face (an essential component of speech-reading [5]), and due to the context-dependent nature of human languages, captions also introduce processing delays when used during conversation, inhibiting conversational flow. Many previously-developed techniques require substantial amounts of training. For example, an evaluation of the iBaldi visualisation technique had participants train for between 30 and 50 hours [7]. Learning to use the existing techniques requires a substantial investment of time on behalf of speech-readers. Obviously, any possible way of reducing training time will benefit speech-readers as they learn to use any new techniques, as well as encourage adoption.

2. SOLUTION

To address the limitations of current techniques, I propose that we visualise the phonemic units of speech. I have developed PhonemeViz, which unlike many existing visualisation techniques presents simple textual representations of a speaker's phonemes to the speech-reader. These textual phoneme descriptors are presented radially off to one side of the speaker's face (to avoid obscurement), and uses movement (essentially a 'phoneme meter') to indicate which phoneme has been uttered by the speaker. Phonemes belonging to the same viseme group are separated and presented in spatially distinct regions of the periphery and alphabetically top to bottom, so the phoneme meter points in distinctly different directions to indicate which phoneme within a viseme group was just uttered by the speaker.

By combining peripheral movement with distinct visual patterns, speech-readers should be able to attend to the speaker's face while being able to disambiguate confusing viseme-to-phoneme mappings, and therefore improve their understanding during conversation.

My doctoral research will be comprised of the following:

1. The design and implementation of the PhonemeViz technique followed by a quantitative evaluation of PhonemeViz against five other visualisation techniques (from related work) along with a no visualisation control condition. Each technique will be superimposed on videos of individual words. The aim of the evaluation is for the participant to use each visualisation technique to determine which word is being spoken.

2. An implementation of PhonemeViz in a VOIP application such as Skype or Google Hangouts, using a speech recognition engine such as PocketSphinx (`cmusphinx.sourceforge.net/`) to recognise the initial phoneme of words spoken by the speaker. To evaluate this technique, participants' accuracy and subjective impressions will be compared against automatically transcribed captions (generated using the speech recognition engine) and a no visualisation control condition.

3. An implementation of PhonemeViz in a wearable device such as Google Glass or the Epson Moverio Glasses. A longitudinal evaluation of my visualisation on this device, where accuracy and subjective impressions will be gathered and compared to current techniques.

3. STAGE OF RESEARCH

Regarding my program of study, I have completed my transfer of ordinance and I am currently in my second year of study. I also took part in the ASSETS 2014 Student Research Competition and was awarded first place in the graduate category [4].

Regarding my PhD research, I have partially completed item 1) and item 2) in the solution list above, and have yet to begin work on item 3). Item 1) was submitted but not accepted to ASSETS 2015. The reviewers provided some constructive feedback on this item of research. I am hoping to review areas of 1), re-run my evaluation and submit it as a full paper to CHI 2016. I have made progress on sections of Item 2) and hope to submit it to ASSETS 2016. Item 3) is still outstanding, but I plan to complete it by the spring of 2016.

4. CONTRIBUTIONS

The primary contribution of this research is a visualisation which improves speech-reading accuracy and reduces workload of the task for a speech-reader.

Secondary contributions include:

1. The general approach of visualising phonemic units to aid with speech comprehension.

2. A speech language model which focuses on recognition of the initial consonant of spoken words, instead of traditional models which aim to recognise whole words or phrases using context.

3. A longitudinal evaluation of my visualisation on a device with a wearable display.

5. DOCTORAL CONSORTIUM OUTCOMES

Attending the ASSETS 2015 Doctoral Consortium will present a number of opportunities for me. First, I will have the opportunity to present my research to senior researchers and fellow students. I appreciate any opportunity I have to describe my research to others.

Second, this presentation will allow me to gain valuable feedback on the approach I have taken in my research to date. This feedback will contribute to my general growth as a researcher, to my production of quality research in the future and to the overall quality of my PhD thesis.

Third, the consortium will present an opportunity to network with senior researchers in this field as well as future colleagues or potential collaborators. These are the individuals I will be working and collaborating with for the remainder of my career. Fourth, potential career paths and opportunities may arise from these discussions.

Finally, the consortium will allow me the opportunity to provide feedback to fellow students regarding their research and presentation styles. I enjoy these opportunities to understand and be made aware of other researchers' work, and it allows me to provide insights that I have gained from my own research.

6. REFERENCES

[1] C. R. Berger and R. J. Calabrese. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *HUM. COMMUN. RES.*, 1(2):99–112, 1975.

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

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

[4] B. M. Gorman. Visaural: A wearable sound-localisation device for people with impaired hearing. In *Proc. ASSETS '14*, pages 337–338. ACM, 2014.

[5] 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.*, 65(4):536–552, 2003.

[6] 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*, pages 265–270, 2004.

[7] 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.

[8] J. E. Shoup. Phonological aspects of speech recognition. *Trends in speech recognition*, pages 125–138, 1980.