Lipreadability of a Synthetic Talking Face in Normal Hearing and Hearing-Impaired Listeners

Catherine Siciliano, Andrew Faulkner, Geoff Williams

Department of Phonetics and Linguistics
University College London
catherine@phon.ucl.ac.uk, andyf@phon.ucl.ac.uk, geoff@phon.ucl.ac.uk

Abstract

The Synface project is developing a synthetic talking face to aid the hearing-impaired in telephone conversation. This report investigates the gain in intelligibility from the synthetic talking head when controlled by hand-annotated speech in both 12 normal hearing (NH) and 13 hearing-impaired (HI) listeners (average hearing loss 86 dB). For NH listeners, audio from everyday sentences was degraded to simulate the information losses that arise in severe-to-profound hearing impairment. For the HI group, audio was filtered to simulate telephone speech. Auditory signals were presented alone, with the synthetic face, and with a video of the original talker. Purely auditory intelligibility was low for the NH group. With the addition of the synthetic face, average intelligibility increased by 22%. The HI group had a large variation in intelligibility in the purely auditory condition. They showed a 22% improvement with the addition of the synthetic face. For both groups, intelligibility with the synthetic face was significantly lower than with the natural face. However, the improvement with the synthetic face is sufficient to be useful in everyday communication. Questionnaire responses from the HI group indicated strong interest in the Synface system.

1. Introduction

For hearing-impaired persons, auditory information is often insufficient for successful communication in the absence of the visual signal. This is particularly relevant for telephone communication, where the hearing-impaired user is at a distinct disadvantage. Recent technological developments have shown that the videophone can be a valuable form of communication for hearing-impaired people, providing essential visual speech information. However, videophones require expensive equipment at both ends and high bandwidth, impracticalities that have led to very limited uptake of the technology. Research has already demonstrated that synthesized visual face movements, driven by an automatic speech recognizer (ASR), can be used to deliver speech information that is unavailable through the auditory channel to hearing-impaired individuals [1, 2]. The goal of the Synface project is to develop a multilingual synthetic talking face that is driven by telephone speech to provide important visual speech information for the hearing-impaired in telephone communication. This technology has the distinct advantage that only the user on the receiving end needs special equipment.

Quality assessment of a system such as Synface is twofold. We must examine both the usefulness of the speech articulations as well as the accuracy of the ASR. In order to maintain near real-time communication, it is necessary that the synthetic face be driven by a real-time recognizer of telephone speech. Such a recognizer will almost certainly produce errors. In this report, we focus on the potential usefulness of the synthetic face when the input is error-free, i.e. from hand-labeled speech.

Previous work on the Teleface project at KTH, Sweden, with an earlier version of the synthetic face designed for the same application showed that both normal hearing and hearing-impaired listeners gained improvement in speech intelligibility with the synthetic face [1, 2]. One goal of the Synface project is to expand the capability of the synthetic face to more European languages, including British English and Dutch. This paper compares previously reported intelligibility studies with normal hearing native British English speakers [3, 4] to new results from tests with hearing-impaired native speakers of British English. Objective feedback about the system was also obtained. Results will be used to define the target group most likely to benefit from the Synface system, and improve weak areas of the system where possible.

2. Visual Speech Synthesis

The talking head used in this study comes from KTH, Sweden. The facial model is implemented as a wire-frame polygon surface that is deformed to simulate speech through a set of parameters. The underlying wireframe model and smooth-rendered surface are shown in Figure 1. Parameters for speech articulations include jaw rotation, labiodental occlusion, lip rounding, bilabial occlusion, tongue tip, tongue length, mouth width and lip protrusion. Control software for the face uses articulatory rules to derive oral parameters. The target values and time offsets of each of the oral parameters for each phone are defined by hand. To synthesize visual speech, the software maps a time-labeled phone sequence to these targets to control the face, using a simple non-linear interpolation scheme to model coarticulatory effects in a smooth and realistic manner. For further details of the implementation of the synthetic face, see [1, 2, 3].
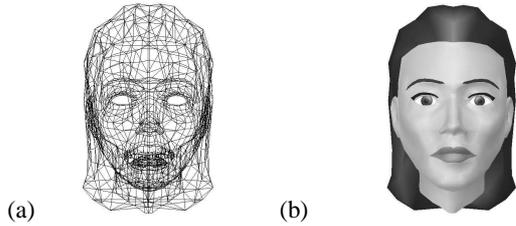
Figure 1: (a) Wireframe model and (b) smooth rendered surface model of the synthetic face.

3. Sentence Intelligibility

3.1 Method

12 normal-hearing (NH) and 13 hearing-impaired (HI) native speakers of British English were tested. Normal hearing subjects were all UCL staff or students. Hearing-impaired subjects volunteered in response to an article in [5]. HI participants had an average hearing loss of 86 dB (range 51-105 dB). During the experiment, the subjects used the hearing aids they normally would use in everyday conversation. The speech material consisted of the BKB sentences, which are used commonly in speech audiometry assessment [6]. No subjects were familiar with the speech material prior to their participation. Each list contains a set number of key words to be used in scoring. Three separate visual conditions were used: audio-only, synthetic face and natural face. The audio-only condition provides a baseline intelligibility level, while the natural face represents the optimum performance achievable in practice.

For the NH listeners, a noise-excited channel vocoder was used to degrade speech, through reconstruction of the signal as a sum of multiple contiguous channels of band-pass filtered noise over a specified frequency range. See [7, 8] for a more detailed description of the technique. In order to simulate a controlled range of hearing impairments, we examined two levels of signal degradation. We determined from a pilot study that 2 and 3 frequency bands were optimal for this experiment, as the conditions provided some auditory information yet forced the listener to rely on the synthetic face. For HI listeners, the audio was bandpass filtered with a 6th order IIR filter from 300 Hz to 3400 kHz. This was done in order to simulate telephone quality speech, the eventual application environment of the Synface system.

The natural face recordings were digitized and then recombined with the filtered audio. Semi-automatic labeling of the audio was performed by means of a forced alignment procedure, and was subsequently hand-corrected. These labels served as input to the facial synthesis software, which mapped the labels to the appropriate facial targets. The facial syntheses were combined with the degraded audio into AVI files. For the audio-only condition, a single image was combined with the audio. All video frame rates were 25 frames per second.

3.2 Procedure

Subjects were seated in front of a computer screen and a loudspeaker (HI, 70 dBA SPL) or headphones (NH), were presented with a sentence, and were then asked to repeat what they perceived. In order to avoid spelling and typing errors across subjects, the experimenter entered the participants' response into the computer. When the experimenter was unclear of the response, the subject was asked to repeat what they had said. Subjects were given practice lists for each visual condition to accustom them to the modified speech signals and the synthetic face. Following practice, each subject heard three (NH) or four (HI) lists in each condition. Presentation of conditions was randomized, with condition held constant during each list.

3.3 Results

The number of keywords identified correctly was counted, ignoring errors of morphology. Scores are expressed as percent of keywords correct. Three of the HI subjects had audio-only scores of greater than 90% keywords correct, and would not be expected to show significant improvements in auditory-visual conditions. These subjects do not fall within the target range of recruited volunteers, and their results were thus excluded from the analyses. Box-and-whisker plots of the results for the HI group and for both conditions for the NH group are given in Figure 2.
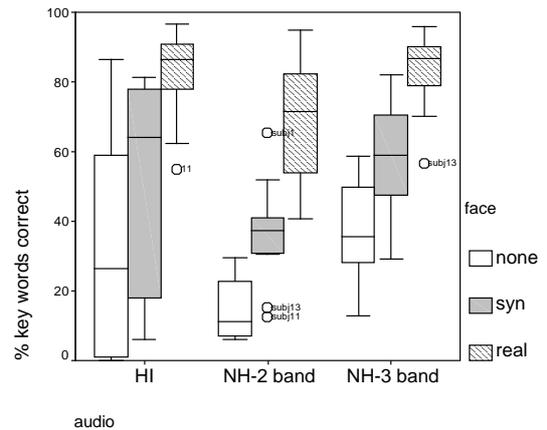

Figure 2. Sentence intelligibility versus facial condition for HI and NH groups.

The mean number of key words identified by each subject in each condition was entered into a repeated-measures ANOVA, with within-subject factors of visual signal (HI) or visual and audio signal (NH). For the HI group, the effect of the visual signal was highly significant ($p < 0.001$). A planned pairwise comparison showed that the presence of a synthetic face led to a significant increase in intelligibility compared to the absence of a face ($p < 0.01$). The difference in intelligibility between the synthetic face and the natural face was highly significant ($p < 0.001$).

Results for the NH group followed a similar pattern. The effects of auditory and visual signals were highly significant ($p < 0.001$), and showed no significant interaction. Planned pairwise comparisons showed that there was a significant increase in intelligibility with the synthetic face compared to

the absence of a face (always with p < 0.001). The natural face provided significantly higher intelligibility than the synthetic face (p < 0.001).

The data show a significant benefit from the synthetic face for both NH and HI groups. Intelligibility on the purely auditory conditions was low (average of 14% for the NH 2-band vocoder, 37% for the NH 3-band vocoder, and 31% for the HI group). Both the HI and NH groups had an average improvement with the synthetic face of 22 keywords out of 100. The magnitude of the intelligibility increase for the synthetic face compared to no face was statistically reliable.

3.4 Discussion

For the HI group, there was a large degree of variability across subjects with both the audio-only and synthetic face conditions. Some subjects scored very low with the audio alone, but showed scores closer to the natural face condition with the synthetic face. However, one subject scored low (<10%) in both the audio-only and synthetic face conditions but scored 78% with the natural face. The NH group did not exhibit such variation. A possible explanation for this is that the HI group had a larger variation in lipreading ability. Another explanation may be that the HI group had different and more varied motivations for using the synthetic face, both within the group and compared to the NH group. Subjective feedback from the HI group indicated this may be the case.

Any systematic account for this variation is helpful in defining the potential target market for the Synface system. In a previous study [2] the authors found that, for speakers of Swedish, the people whose intelligibility with the audio-only condition fell within the 40-80% range had the largest improvement with the synthetic face. A further analysis with scatter plots of the HI results was conducted to determine whether this was the case for this group. Scatter plots of baseline intelligibility versus improvement with the synthetic face and hearing loss versus improvement with the synthetic face are given in Figure 3. In Figure 3a, the plot indicates there is no significant correlation between percent improvement with the synthetic face and hearing loss. This was confirmed by a linear regression of the data. This may indicate variation in the quality of the participants' hearing aids and their ability to use them, or simply reflect large variations in lipreading ability. A linear regression of percent correct with audio-only scores with hearing loss showed no significant correlation, suggesting that average hearing loss is not a strong predictor of who will benefit from the synthetic face.

The plot of improvement with synthetic face versus intelligibility in the audio-only conditions shows that, excluding cases where intelligibility in the audio-only condition was near zero, there is a correlation between the two variables. A linear regression showed that this correlation is significant (p < 0.01). Contrary to what was reported in [2], it was the subjects with low audio-only intelligibility (as low as 10%) who gained the largest amount of information from the synthetic face: this may define the users who would gain most from a synthetic face in everyday telephone communication. Participants who scored less than 10% on the audio-only condition also showed gains in intelligibility with the synthetic face. However, the improvements were small and unlikely to be useful in a telephone conversation. These results suggest that the target group of users may be broader than initially suggested in [2], at least for speakers of British English.
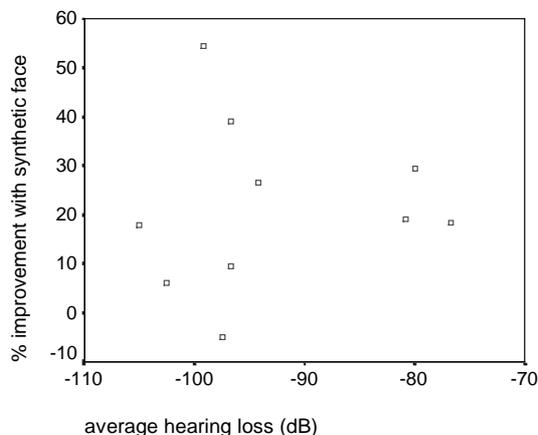
(a)
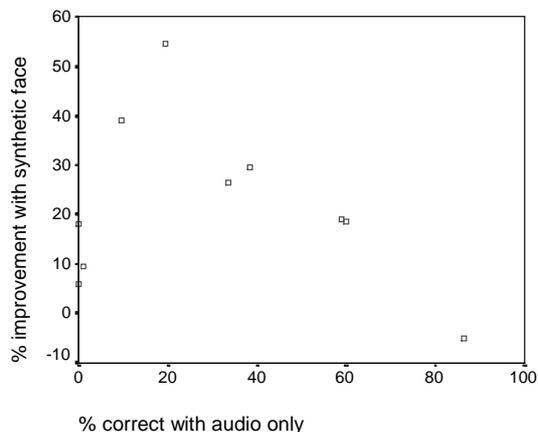
(b)

Figure 3. *(a)* Average hearing loss versus percent improvement with synthetic face. *(b)* Percent correct with audio-only versus improvement with synthetic face.

The improvements seen here are larger than those reported in [2]. This may be due to language differences and improvements in the facial synthesis. Additionally, in [2] one condition was a synthetic face driven by an automatic speech recognizer, which would be likely to lead to lower intelligibility.

4. Questionnaire

In addition to the intelligibility tests, the HI volunteers were asked to complete a subjective questionnaire on the usefulness of the Synface system. One phase of the Synface project will involve a large-scale subjective evaluation of a prototype system. Therefore, the questionnaire here was not meant to be comprehensive, but rather to serve roughly as a preliminary assessment by the potential market. To give the participants a better idea of what the eventual product might look like, they were shown a display of a prototype system before completing the questionnaire.

When asked how useful they thought the system was, all but one HI participant reported that the system was useful or very useful. The majority of participants indicated they would prefer a Synface system to a telephone relay service. Most expressed interest in using the Synface system at home.

HI participants were also given the opportunity to respond freely with their good and bad points about the Synface system. Several noted the advantage of Synface that there is no third party required in communication. However, most subjects expressed an interest in adding emotion to the synthetic face.

5. Conclusions and Future Work

The experiments presented here indicate that the synthetic face in its current form can be used to transmit important visual speech information. Since intelligibility with the synthetic face falls short of that with a natural face, though, there is room for improvement. Results from more analytic segments are being used to improve the face synthesis through frame-by-frame comparisons with the natural face. Furthermore, current work on the project is examining whether a data-driven, rather than rule-driven, method of visual speech synthesis yields greater intelligibility.

While a large-scale assessment of a prototype system in the homes of actual users is planned for a later stage in the project, the results here will be helpful as a preliminary analysis of a potential market. Subjective feedback obtained from volunteers indicates that there is interest among the hearing-impaired community in the Synface system. However, several participants expressed strong interest in adding emotion to the synthetic face. Research into emotion in multimodal speech synthesis is being carried out in parallel with the project, and may eventually be incorporated into the system.

Since the synthetic face is intended to be driven by an automatic speech recognizer rather than hand-labeled speech, the next phase of intelligibility experiments will involve examining the effects of ASR errors on speech perception. Work has begun to examine both the individual and combined effects of timing errors and segmental errors on intelligibility. Further experiments are planned to simulate offline a working Synface system. Results from these studies will be used to define the minimal requirements of the automatic speech recognition software.

6. Acknowledgments

7. References

[1] Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., Öhman, T., "The Teleface project: Multimodal speech communication for the hearing-impaired", *Proceedings of Eurospeech*, Rhodes, Greece, 1997.

[2] Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M, Spens, K.-E., Öhman, T. "Synthetic talking faces as a lipreading support", *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[3] Siciliano, C., Williams, G., Beskow, J.., Faulkner, A. "Evaluation of a synthetic talking face as a communication aid for the hearing impaired", *Speech, Hearing and Language: Work in Progress*, **14**: 51-61, 2002. http://www.phon.ucl.ac.uk/home/shl14/pdffiles/sicilianoWBF.pdf

[4] Siciliano, C., Williams, G., Beskow, J. and Faulkner, A., "Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired", *15th International Congress of the Phonetic Sciences*, Barcelona, Spain, 2003.

[5] Nakisa, M., "Talking heads", *One in Seven*, RNID, Issue 31, 2002.

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

[7] Villchur, E., "Electronic models to simulate the effect of sensory distortions on speech perception by the deaf", *J. Acoust. Soc. Amer.* **62**: 665-674, 1977.

[8] Shannon, R. V., Zeng, F-G., Kamath, V., Wygonski, J. and Ekelid, M., "Speech recognition with primarily temporal cues", *Science*, **270**, 303-304, 1995.