PAPER 6.

Siciliano, C., Williams, G., Beskow, J. and Faulkner, A. (submitted2). Evaluation of a Multilingual Synthetic Talking Face as a Communication Aid for the Hearing Impaired, to appear in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03)*, Barcelona, Spain.

EVALUATION OF A MULTILINGUAL SYNTHETIC TALKING FACE AS A COMMUNICATION AID FOR THE HEARING IMPAIRED

Catherine Siciliano†, Geoff Williams†, Jonas Beskow‡ and Andrew Faulkner†

† Department of Phonetics and Linguistics, University College London, United Kingdom
‡ Centre for Speech Technology, KTH, Stockholm, Sweden

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. Audio from Swedish, English and Dutch sentences was degraded to simulate the information losses that arise in severe-to-profound hearing impairment. 12 normal-hearing native speakers for each language took part. Auditory signals were presented alone, with the synthetic face, and with a video of the original talker. Purely auditory intelligibility was low. With the addition of the synthetic face, average intelligibility increased by 20%. Scores with the synthetic face were significantly lower than for the natural face for English and Dutch, but not Swedish. Visual identification of English consonants showed that the synthetic face fell short of a natural face on both place and manner of articulation. This information will be used to improve the synthesis.

1. INTRODUCTION

For the hearing impaired community, 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, can be used to deliver phonetic 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 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.

This paper presents results from multilingual perceptual studies assessing the intelligibility of a synthetic talking face driven by hand-annotated speech, in order to define the potential usefulness of the talking face, and to determine areas for improvement in the synthesis. The first experiment consisted of a series of intelligibility tests with native listeners of Swedish, English and Dutch using degraded speech. The use of degraded speech signals with low inherent intelligibility forces the listener to rely more heavily on the visual speech information so that a better understanding of the utility of the synthetic face is

gained. The degradation simulates in normal hearing listeners the reduced intelligibility seen in severe-to-profound hearing impairment by the reduction of spectral detail. A second experiment compared the intelligibility of English VCV segments between a natural face and the synthetic face, with no audio signal, in order to pinpoint the strengths and weaknesses of the visual face synthesis.

2. VISUAL SPEECH SYNTHESIS

The talking head used in this study comes from KTH, Stockholm. The facial model is implemented as a wire-frame polygon surface that is deformed to simulate speech through a set of parameters. 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. For each language, 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].

3. SENTENCE INTELLIGIBILITY

3.1 Method

3.1.1 Subjects

17 female and 19 male normal hearing subjects were tested, 12 from each of the language groups. Subjects were all university or institution staff and students. All were native speakers of the relevant language. No subjects were familiar with the speech material prior to their participation.

3.1.2 Speech material

Recordings of native speakers of a standard dialect of the respective language were obtained. Speakers were two females (English and Dutch) and one male (Swedish). The speech material consisted of lists of everyday sentences in Swedish, English and Dutch designed to test speech perceptual ability. All were similar in complexity and vocabulary. Each sentence contained a set number of key words to be used in scoring. A more detailed description of the material is given in [4, 5 and 6].[1]

3.1.3 Stimuli

Three separate visual conditions were used: audio-only, synthetic face and natural face. Each of these was combined with degraded audio. The audio-only condition provides a baseline intelligibility level for the degraded speech, while the natural face represents the optimum performance achievable in practice. We used two levels of audio degradation, which allowed us to study the usefulness of the synthetic face over a range of simulated hearing loss.

A noise-excited channel vocoder was used to reconstruct the speech signal as a sum of multiple contiguous channels of white noise over a specified frequency range [7]. This processing reduces the spectral detail of speech to that represented by the relative amplitudes in adjacent frequency bands covering equal basilar-membrane distances and

[1] Material in [6] was translated into Swedish from English.

spanning the frequency range 100-5000 Hz, whilst preserving temporal envelope information within each band up to modulation rates of 16 Hz. Amplitude envelopes from each analysis band were extracted by half-wave rectification and smoothing, and used to modulate the amplitude of white noise that was subsequently band-pass filtered with a filter matching the analysis filter for that band. The final signal was the sum of the modulated and band-pass filtered noises. A pilot study indicated that a 4-band vocoder yielded auditory intelligibility approaching that with unprocessed speech, and was thus unlikely to show reliable improvements in intelligibility between visual conditions. Therefore, we used 2 and 3 frequency band vocoders in the main experiment.

The natural face recordings were digitized and then recombined with the degraded audio. Semi-automatic labeling of the audio was performed by means of a forced alignment procedure, which was later hand-corrected. The labels were used to drive the facial synthesis. For the audio only condition, a single image was combined with the audio. All frame rates were 25 Hz.

3.2 Procedure

Subjects were seated in front of a computer screen and a loudspeaker or headphones, were presented with sentences in their native language, and were then asked to repeat what they perceived. The test leader noted the subjects' response. Subjects were given practice lists for each visual condition to accustom them to the modified speech signals. Following practice, each subject heard either two or three lists in each condition. Presentation of conditions was randomized, with each list comprising a single condition. Due to the amount of English and Swedish material, the test was run over two sessions.

3.3 Results

The number of keywords identified correctly was counted, ignoring errors of morphology. Scores are expressed as percent of keywords correct. Box-and-whisker plots of the overall results and for each of the three languages are shown in Figure 1.

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 auditory signal and visual input, and the between-subject factor of language. Within each language group, the effects of auditory and visual signals remained highly significant ($p<0.001$), and showed no significant interaction. Planned pairwise comparisons showed that for each language group, the presence of a synthetic face led to a significant increase in intelligibility compared to the absence of a face (always with $p<0.001$). For the Dutch and English groups, the natural face provided significantly higher intelligibility than the synthetic face ($p \alpha 0.001$). For the Swedish group, however, this difference was not significant.

The data show a significant benefit from the synthetic face under the degraded auditory conditions. Intelligibility on the purely auditory conditions was low (average of 7% for the 2-band vocoder and 30% for the 3-band vocoder) and representative of intelligibility in this target group for the same or similar sentences. With an average improvement of 20 words out of 100 (range 13.6 to 27.5), the magnitude of the intelligibility increase for the synthetic face compared to no face was broadly consistent, statistically reliable, and large enough to be important in everyday communication.

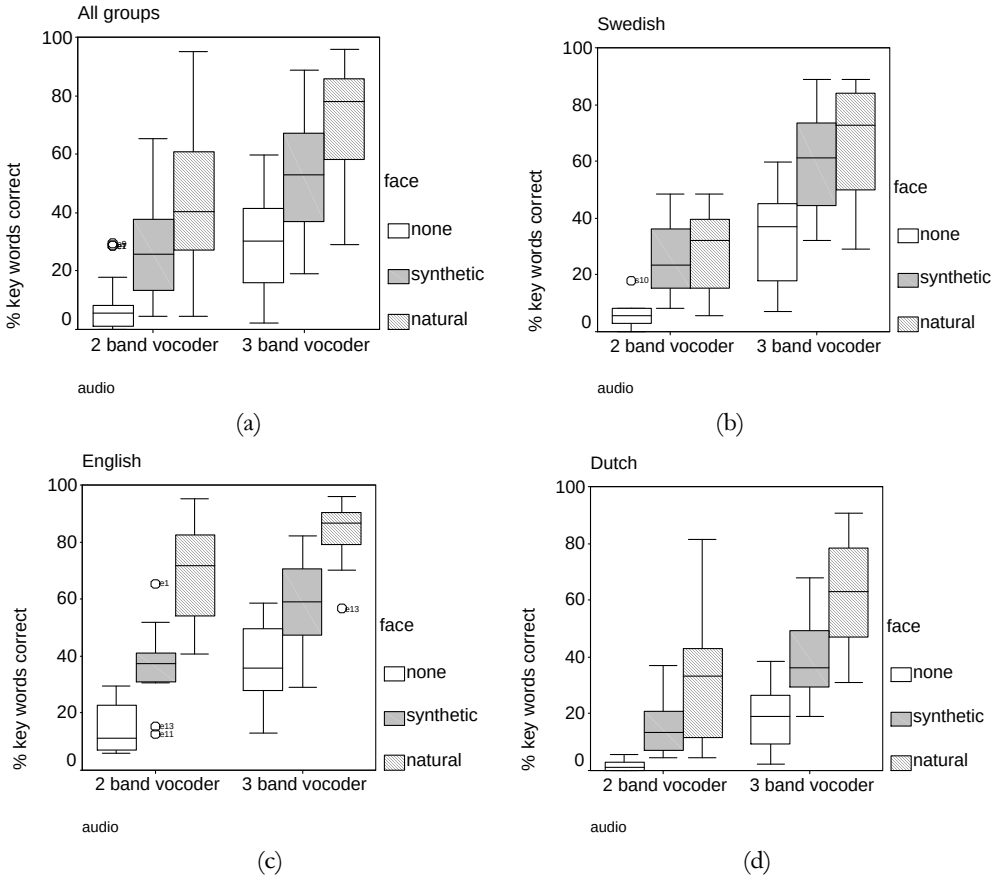

Figure 1. Sentence intelligibility for degraded speech with synthetic and natural visual speech. (a) All three language groups (b) Swedish (c) English and (d) Dutch.

The degree to which the synthetic face fell short of the advantage from the natural face was more variable, with an average of 18 words, but a range of 14.7 to 31.7 words. This variability is primarily due to the very small differences (of around 5 words) between the synthetic and natural faces for the Swedish group. This could be a result of the synthetic face functioning better in the language for which it was primarily developed, or alternatively because the Swedish speaker was relatively difficult to speech-read. The spread of scores for the Swedish face in the 3-band vocoder condition is very large compared to that for the natural English face, which does suggest that the Swedish talker is relatively difficult to speech-read, at least for some of the test subjects. This could be due to language differences, gender differences, or general variability in talker intelligibility.

4. VCV INTELLIGIBILITY

Since the information transmission from the synthetic face does not approach that from the natural face in the English or Dutch subject groups, it is likely that further refinements of the face synthesis in these languages would be advantageous. We therefore conducted experiments comparing the intelligibility of more analytic VCV segments between a natural face and the synthetic face.

4.1 Method

4.1.1 Subjects

Subjects were 5 male and 5 female native speakers of British English, all with normal hearing. All were university staff or students.

4.1.2 Stimuli

24 English consonants in left and right /i/, /a/, and /u/ contexts were used as stimuli. The consonants were /p b m t d n s z l r w j ʒ ʃ tʃ dʒ f v k g ŋ h θ ð/. Recordings of two female native British English speakers were used. Label files to drive the face synthesis were generated from the original audio in the same manner as in 3.1.3 above. There were 3 tokens of each VCV for each speaker, and corresponding tokens for each of these with the synthetic face. In order to derive maximal information about the visual signal, no audio signal was presented.

4.2 Procedure

Subjects were shown a movie of a VCV token, and then asked to indicate the correct consonant in a forced-choice task. A subset of VCV segments, one for each consonant, was given initially as practice.

4.3 Results

Accuracy was low overall, with 13.6% correct responses for the synthetic face and 23.4% for the natural faces. The difference in intelligibility between the synthetic and natural faces was highly significant ($p \alpha 0.001$). Feature analysis using information transfer measures is summarized in Table 1. The analysis indicates that the synthetic face falls short of the natural faces for both place and manner of articulation. However, information transfer for manner of articulation is much lower than place information transfer in general. The voicing feature contributes negligibly. Therefore, information about any systematic confusions in the data is more likely to be obtained from analyzing the place feature.

Figure 2 shows bubble plots of the place feature. The plots demonstrate that intelligibility of the synthetic face approaches that of the natural face for frontal places of articulation, such as bilabials and labiodentals. However, the synthetic face falls short of the natural face for back articulations. The plots indicate a large proportion of alveolar responses for the synthetic face compared to the natural face. The raw confusion matrices in fact reveal a large number of /l/ responses for the synthetic face compared to the natural face. This was also reported anecdotally by a few subjects. Several alveolar consonants were defined in the synthesis software to have identical articulations, which resemble /l/, accounting for a portion of the alveolar responses. For palatals and dentals, also confused as alveolars, the synthetic face must lack distinctive visual information necessary to recognize these consonants. In improving the synthetic face, then, we will need to incorporate more detail about place and manner of articulation.

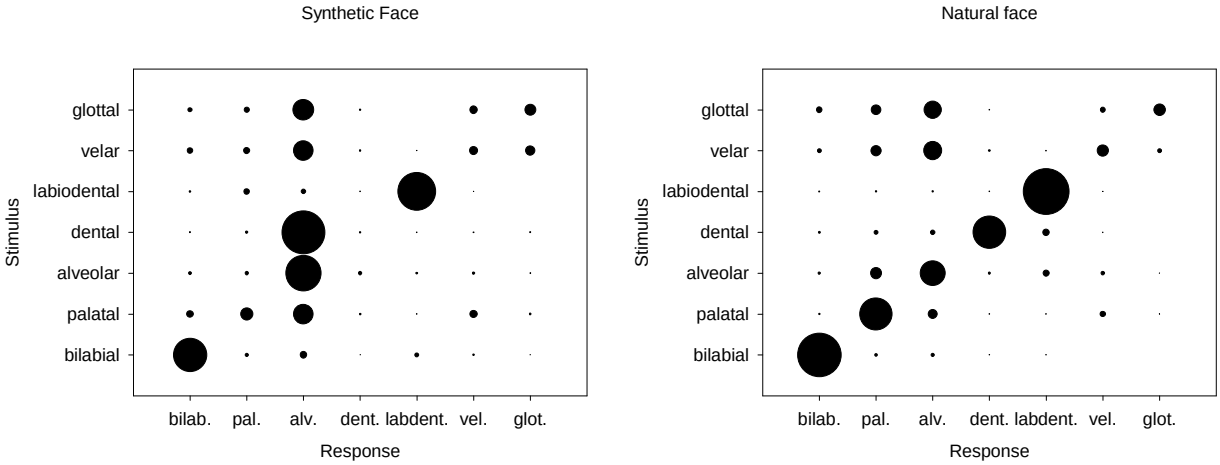

Figure 2. Place confusions for the (a) synthetic face and (b) natural faces. Area of circle represents scaled proportion of responses. Largest bubble = .878.

FEATURE	%TRANS		%CORRECT	
	Synth	**Real**	**Synth**	**Real**
place	18.6	31.3	44.9	57.2
manner	2.9	11.5	32.9	47.5
voicing	0.3	0.8	57.5	56.8

Table 1. Information transfer and percent correct of features for real vs. synthetic face

CONCLUSIONS AND FUTURE WORK

The two experiments suggest that the synthetic face in its current form can be used to transmit important visual phonetic information. However, as intelligibility with the synthetic face falls short of the natural faces, there is room for improvement in the face synthesis. We eventually intend to employ a data-driven method of visual speech synthesis. In the mean time, results from intelligibility tests with analytic VCV experiments will be used to further improve the face synthesis. This will be accomplished through frame-by-frame comparisons of the synthetic face with the natural faces. Work is also in progress to replicate these findings with hearing impaired users, and to determine the effects of errors in automatic speech recognition on synthetic face intelligibility.

ACKNOWLEDGMENTS

The Synface project is funded by the European Commission grant IST-2001-33327. We are grateful to Eline Rikken and Nic van Son of Viataal, the Netherlands, and Eva Agelfors and Inger Karlsson of KTH, Sweden, for their help in carrying out these experiments.

REFERENCES

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

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

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

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

[5] A. MacLeod and Q. Summerfield. "A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use", *British Journal of Audiology*, **24**, pp. 29-43, 1990.

[6] R. Plomp and A. Mimpen. "Improving the reliability of testing the speech-reception threshold for sentences", *Audiology*, **18**, pp. 43-52, 1979.

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