DIFFERENCES IN SPEECHREADING A SYNTHETIC AND A NATURAL FACE

Tobias Öhman and Magnus Lundeberg
Centre for Speech Technology, KTH, Sweden

ABSTRACT

In this paper we describe a method to improve speech movements of synthetic faces. Intelligibility tests were used to find phonemes or groups of phonemes where a natural face was significantly better in transmitting audio-visual information compared to a synthetic face. Optical measurements of the speech movements of the human face were performed, and several visual features were identified which could account for the intelligibility differences. As an application of this method, we adjusted our visual speech synthesis according to the results. In a new audio-visual intelligibility test, we could confirm the improvements of the synthetic face.

1. INTRODUCTION

Speech has visual as well as auditory components, and the perceiver makes use of both. In this sense, then, the perceiver does not distinguish between visual and acoustic cues; he/she uses whatever is available in the signal to make as much sense as possible of the message, from whichever sensory mode the information originates. Depending on the situation, different cues are used to different degrees. Relevant to our study is the fact that in acoustically noisy environments and for hearing-impaired people, features originating from the visual domain play a more prominent role [1]. In agreement with this view, we analyse confusions from audio-visual stimuli, and our efforts to improve the visual synthesis are based on an audio-visual foundation. In other words, we want to improve the articulations of the artificial face mainly where it is needed for speechreading, taking into consideration that auditory information is simultaneously available to a perceiver. It is the end result that counts.

The concept of the viseme is used in many different ways in the literature. There are at least two interpretations to keep separate. In the first, the viseme is, as the name *vis*ual phon*eme* indicates, ascribed the same linguistic/psychological status as the phoneme. This way of using the term should be opposed to a more pragmatic definition, according to which the viseme is a group of visually similar phonemes. We reject the first definition, and in the rest of this paper the term viseme refers to a group of phonemes which are visually closer to each other than to other phonemes. This is not to say that members of the same viseme class are visually indistinguishable in all contexts.

The workbench for our study is the rule-based KTH audio-visual text-to-speech-synthesis framework [2]. Synthetic faces have been developed and used as human-machine interfaces in applications, such as the spoken dialogue systems Waxholm and August [3][4]. In another application oriented project, the Teleface project [5], we are incorporating a synthetic face as a speechreading aid for the hard of hearing in telephone communication. We found a considerable increase in the intelligibility of telephone speech when visual information provided by our synthetic face was added. We also showed that a natural face, instead of a synthetic face, was an even better speechreading support. This observation constitutes the starting point for our current work. The question at issue is how we can use the results from the intelligibility tests, in combination with measurements of the speech movements of the natural face, to improve the visual speech synthesis.

2. PREVIOUS INTELLIGIBILITY TESTS

As previously reported [6], we have performed audio-visual intelligibility tests where a natural voice was presented together with either an animated synthetic face or video sequences of a natural face. Figure 1 shows the synthetic and the natural face.

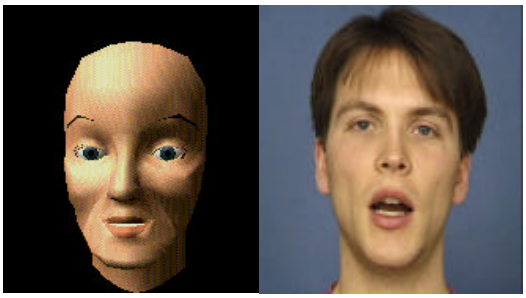


Figure 1. The synthetic and the natural face

The synthetic visual speech stimuli for the tests were generated using the audio-visual speech synthesis on speech material that had been phonetically labelled using forced alignment and manual fine-tuning. Eight synthesis parameters (such as bilabial occlusion, rounding etc.) are used to control the articulation of the face. All parameters have specified target values for the vowel visemes. For each consonantal viseme, a subset of the parameters has specified target values, and values for the remaining parameters are obtained by interpolating the values from the adjacent vowels.

The speech material was VCV-syllables, with the consonants /b, d, g, p, t, k, s, ʂ, ç, f, v, m, n, ŋ, j, l, r/ symmetrically surrounded by the vowels /a/ and /ʊ/. Each stimulus was presented once to each of the eighteen normal hearing subjects and nine hearing impaired subjects. The audio for the normal hearing subjects was degraded by white Gaussian noise, 3 dB SNR, to simulate a hearing loss. The result for all subjects was that the video recording of the natural human face was slightly better on average (70 % consonants correct) than our synthetic face (65% correct). This points to a considerable improvement in intelligibility compared to presenting the auditory signal alone (52 % correct).

3. HYPOTHESIS TESTING

We used four confusion matrices from the intelligibility tests: two for the natural face (one per vowel context) and two for the synthetic face. In this section, we propose an algorithm to point out the most prominent confusions and determining how significant they are. Given that a significant difference between the faces is found for some distinction, a second issue is to make use of the optical measurements to explain this difference. In other words, we want to find visual features that are not (sufficiently) utilised in the synthetic face for making the desired distinction.

3.1. Binomial Test

Let N be the sum of corresponding cells in the confusion matrices from the two faces, i.e.,

$$N = c_{ij,n} + c_{ij,s}, \quad (1)$$

where c_{ij} is the number of times when a stimulus of unit *i* was perceived as unit *j*, for the natural and synthetic face, indicated by the indices *n* and *s*, respectively. For example $c_{nl,s}$ is the number of times /n/ was perceived as /l/ for the synthetic face. The number of confusions from one of the faces, $c_{ij,f}$, can be seen as the outcome of a stochastic variable, $C_{ij,f}$. Since either the natural face or the synthetic face constituted the stimuli (together with the natural voice, as described in section 2), $C_{ij,f}$ has a binomial distribution, with the parameters *p* and *N*. Here *p* is the *conditional* probability that face *f* constituted the stimulus, *given* that the total number of confusions of this sort were *N*. The probability function is then

$$\Pr(C_{ij,f} = x) = \binom{N}{x} p^x (1-p)^{(N-x)}. \quad (2)$$

Our null hypothesis, $\mathbf{H_0}$, is that the confusion arises with equal probability when either of the two faces constituted the stimulus, (i.e. no face is better than the other). Our alternative hypothesis, $\mathbf{H_1}$, is that the face for which the observed confusions occurred with a lower frequency, was used with a lower probability. In mathematical terms, this is stated as

$$\mathbf{H_0}: p = \tfrac{1}{2}, \text{ and } \mathbf{H_1}: p < \tfrac{1}{2}, \quad (3)$$

where *p* is the probability that the face with the lowest number of observed confusions was used.

Since we will consider many tests (one for each cell in the confusion matrix) it is convenient to report the results in terms of *critical* (significance) levels instead of significance levels. The critical level, α^*, is defined as the smallest significance level at which the null hypothesis would be rejected for the observed statistics. In other words, it is the probability under $\mathbf{H_0}$ that the outcome of C_{ij} is the observed value or lower (since the test is one-tailed). The critical level under $\mathbf{H_0}$ for perceiving stimuli of unit *i* as unit *j*, is then:

$$a^*_{ij} = \sum_{x=0}^{\min(c_{ij,n}, c_{ij,s})} \binom{N}{x} (1/2)^N. \quad (4)$$

This is the value we will report for different confusions in the next section.

4. RESULTS

We have applied the binomial tests to the confusion matrices obtained in the intelligibility tests described in section 2. The results are presented separately for the two surrounding vowels /ʊ/ and /a/. We present the results for the visemes currently used by the face synthesis, and discuss selected confusions between individual phonemes.

4.1. Visemes

The results are presented as matrices of critical significance levels, with the rows corresponding to the stimuli, and the columns to the responses. The matrices should be interpreted in the following way: Positive entries, outside the diagonal, show the critical level for a test where the alternative hypothesis is that the natural face was less likely to be involved in the confusion. Entries preceded by a minus sign are the critical level of significance for tests where more confusions were found for the natural face. In these cases the alternative hypothesis states that the synthetic (not the natural) face was less likely to have been presented. The diagonal (values in bold) corresponds to correctly classified units. A significant difference in this case means that the face with the fewest observations is less likely to have been used when the unit was correctly classified, i.e. has a lower overall performance than the other face.

To keep the faces apart, we have inserted a minus sign before entries corresponding to tests where the synthetic face is in favour, irrespective of whether we are testing for a confusion between units or for a correctly classified unit.

Table 1 shows the results for the consonantal visemes with /a/ as the surrounding vowel. Entries with the value 0.5 (corresponding to exactly equal numbers of confusions for the two faces), or zero (a confusion that never occurred) have been deleted, to make the matrix easier to interpret. Interesting results (with a critical level of significance lower than 10%) have been marked with a grey background.

	-0,4370							
		0,0938	0,1325	0,0010			-0,1875	
		0,3125	**-0,3593**				-0,1481	
			0,2500					
					-0,4606			
		-0,1077	0,2500		0,1250	**-0,2539**		
			0,1189			-0,1094	**0,4439**	
			0,3438			-0,3438		

Table 1. Critical levels of significance in the /a/-context

For the /a/-context, the most definite difference is the confusion from /t, d, n/ to /r, l/. For /t, d, n/ we also have the lowest critical level in overall performance. In both cases, the natural face was

better. In Table 2, the corresponding results are shown for the /ʊ/-context. If we focus on the same viseme, /t, d, n/, we can see that it has been perceived considerably more often as the visemes /k, g, ŋ/, and /r, l/ for the synthetic face. The most prominent diagonal element is also in this case that for /t, d, n/, in favour of the natural face.

	p,b,m	t,d,n	k,g,ŋ	r,l	f,v	s	j,ç	ʂ
p,b,m	**0,3691**		0,1094				-0,2500	
t,d,n		**0,0048**	0,0057	0,0207	-0,3125	0,1250	-0,1917	
k,g,ŋ	0,2500	-0,0178	**-0,1514**					-0,2500
r,l		-0,3438	0,3872	**0,3957**				
f,v	0,1250		0,2500		**0,3822**			
s		-0,3318	0,0078					-0,3770
j,ç		-0,1133	0,0020	-0,2266		-0,3125	**0,4415**	0,3556
ʂ			0,1250			-0,2539		

Table 2. Critical levels of significance in the /ʊ/- context

An interesting result to be noted here is that for the confusion from /k, g, ŋ/ to /t, d, n/. This is an encouraging result, indicating the synthetic face to better in this case. The overall performance for /k, g, ŋ/ points in the same direction, although not as distinct ($\alpha^* < 0.1514$).

4.2. Individual Phonemes

The results for the individual phonemes will only briefly be mentioned here. The overall picture can be summarised in a few words, since most of the entries are zero or near 0.5. For the /a/-context, there is only one confusion worth mentioning. This is when /n/ is perceived as /l/. Here we see a very clear difference between the faces, with a critical level of significance that is less than 0.0010. In the /ʊ/-context, the difference between the faces is prominent in the confusions /n/ to /l/, and /n/ to /ŋ/ (with α^* less than 0.0195 and 0.0352, respectively).

5. VISUAL SPEECH MEASUREMENTS

In the previous section, we saw that there was a difference between the faces in the articulation of /t, d, n/. In this section, we want to specify this difference and find (some of the) responsible visual features, by analysing lip and jaw movements of the natural face for the current pairs of visemes and phonemes. To that end, we used a database of natural speech, which will be described briefly in the following subsection. For further details, see [7]. We then describe some of the visual features relevant to this study, followed by a subsection where hypotheses about the features are tested.

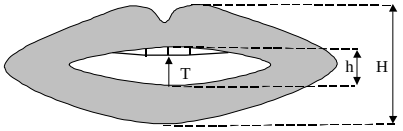
5.1. The Database

The audio-visual speech database consists of video sequences of the same speaker as in the intelligibility tests. 269 sentences and 153 VCV-syllables, including those used in the intelligibility tests, were recorded. The audio tracks of all 422 utterances were phonetically segmented and labelled. Optical measurements of nearly 27.000 video frames (25 frames per second) were automatically performed by an image analysis algorithm. Quantities such as areas, perimeters and positions of articulators were simultaneously measured in the frontal as well as in the lateral view of the face. The resolution was 0.5 mm in both the x and y directions.

5.2. Audio-visual Speech Features

In this paper, we will restrict ourselves to consider only the VCV-part of the database, and only a subset of the measured features. The three features discussed in this study are: internal vertical lip opening, **h**, height of the external lip contour, **H**, and vertical lip translation relative to the upper teeth, **T**. The features are shown in figure 2.

Figure 2. Measured visual speech features



5.3. Hypothesis Testing

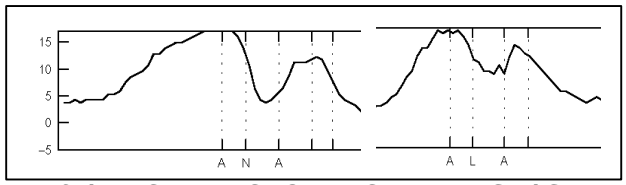
The measured visual features are considered to be normally distributed, with different parameters (mean and standard deviation) for different phonemes or visemes. When testing whether a certain visual feature differs for a pair of phonemes or visemes, a two-tailed, two-sample Student's test, with unequal variance, was used. Our null hypothesis was that the underlying distributions for the two samples have the same mean. For all the discussed distinctions between pairs of phonemes and visemes, and for both vowel contexts, we found significant differences in several measured features. The critical level of significance was well below 10%, in many cases also well below 0.1%.

If we concentrate on the distinction between the phonemes /n/ and /l/, our observations may be summarised in the following two points:

- There is a dip in the trajectories of all the discussed features (**h**, **H**, and **T**), which is deeper for /n/ than for /l/ (Figure 3).
- The difference between /n/ and /l/ in **h** and **H** is roughly the same as the difference in **T**.

From this we conclude that, for /n/, the lower lip is raised a small amount compared to the articulation of /l/, and that the upper lip is unaltered.

Figure 3. Trajectories of the vertical translation feature, **T** (mm)



6. ADJUSTMENTS TO THE SYNTHETIC FACE

We have found differences in the intelligibility of natural and synthetic faces that are both significant and relevant. Especially for the articulation of /t, d, n/, the natural face does a better job, since this viseme is more often perceived as /r, l/ for the synthetic face. This difference can, to a large extent, be ascribed to the

misperception of /l/ as /n/. The measurements show that there is a significant difference between these phonemes, in terms of features related to the vertical distance between the lips. A feature that seems to be responsible for this is the vertical distance between the upper front teeth and the uppermost point of the lower lip (**T**).

As an example, and to see if our ideas are fruitful, we will here apply the results to our visual speech synthesis. We restrict ourselves to consider only the distinction between /n/ and /l/, since our results suggested that there are improvements to be made here.

6.1. Adjusting the Synthesis

According to the results in the previous section, we adjusted our visual speech synthesis for the articulation of the viseme /t, d, n/. In particular, we wanted to improve the distinction between this viseme and /r, l/. According to the measurements, raising the lower lip in the production of /n/ might bring out such a distinction. This is a feature that has yet not been utilised in the synthesis for this distinction. The synthesis parameters that affect the motion of the lower lip are the jaw rotation, bilabial occlusion, and labiodental occlusion. Since a target value for the jaw rotation is specified only for /t, d, n/, and not for /r, l/, and since the bilabial occlusion involves a simultaneous, and unwanted downward movement of the upper lip, we are left with the labiodental occlusion. Since this parameter also most resembles the measured feature, we chose to manipulate only this parameter.

6.2. New Intelligibility Tests

To see whether this adjustment really resulted in the desired improvement, we tested our visual speech synthesis in a new audio-visual intelligibility test. This time we let 31 normal hearing subjects hear degraded audio (natural voice and white noise at 3 dB SNR) and see the two versions of the synthetic face. The speech material (VCV-syllables with C = /l, r, d, n, g, ŋ/ and V = /ʊ, a/), was presented randomly for the two faces. The subjects were asked to respond by reporting which of the consonants they perceived (forced choice). The confusion matrices for the phonemes were then analysed by the method described above (3.1) to compare the performance between the old and the new synthetic face. We found a significant ($\alpha^* < 0.0925$) improvement in the confusion from /n/ to /l/ in the /a/-context, but not in the /ʊ/-context. This was the only significant difference between the faces, both in the /a/- and the /ʊ/-context, and for both phonemes and visemes ($\alpha^* < 10\%$).

7. DISCUSSION

In this paper we have discussed a method to compare confusion matrices originating from two different types of stimuli. In our case, two faces are compared to each other to find out whether, and for what units of speech, one face is more intelligible than the other. In particular, we have discussed how visual speech synthesis can be improved by applying this method to find out where a natural face is better. Measurements of the natural face for the corresponding distinctions may reveal visual features (partly) responsible for this difference, especially if these features are not used by the synthesis to make this distinction.

One issue of statistical importance regards the significance for the whole confusion matrices. A critical significance level can be seen as the expected fraction of tests where the null hypothesis is erroneously rejected. Therefore, since we have performed several significance tests in our study, the significance for the whole test is lower than the individually reported critical levels would indicate.

As an example of this approach, we applied our findings to the synthetic face for the distinction between /n/ and /l/; a distinction where the natural face was found to be better. Measurements guided us in the choice of synthesis parameter to adjust. We increased the labiodental occlusion for the articulation of /t, d, n/, and performed new intelligibility tests. The expected improvement in the distinction between /n/ and /l/ was confirmed for the /a/-context, but not for the /ʊ/-context. Negative side effects could not be found for the tested speech material.

In some cases, the synthesis was found to be significantly better than the natural face. For visemes, /k, g, ŋ/ was significantly more often perceived as /t, d, n/ for the natural face than for the synthetic in the /ʊ/-context ($\alpha^* < 0.0048$). We believe this can be explained by the fact that the tongue is easier to see in the synthetic face for rounded articulations, because it has a better illumination of the oral cavity than the natural face has in our recording.

This leads us to an important note to make, namely that synthetic faces have the potential of becoming an even better speechreading aid than natural a face.

ACKNOWLEDGEMENTS

This work was partly funded by KFB, the Swedish Transport and Communications Research Board.

REFERENCES

[1] Sumby, W.H. and Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26: 212-215.

[2] Beskow, J. 1995. Rule-based visual speech synthesis. In *Proceedings of Eurospeech'95*, Madrid, Spain, 2: 299-302.

[3] Bertenstam, J., Beskow, J., Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa-Leitao, A., Nord, L. and Ström, N. 1995. The Waxholm system – a progress report. In *Proceedings of Workshop on Spoken Dialogue Systems*. Vigsø, Denmark.

[4] Gustafson, J., Lindberg, N., Lundeberg, M. and Svensson, E.-L. 1999. The August spoken dialogue system. To be published in *Proceedings of Eurospeech'99*. Budapest, Hungary.

[5] Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E. and Öhman, T. Synthetic faces as a lipreading support. In *Proceedings of ICSLP'98*. Sydney, Australia.

[6] Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E. and Öhman, T. 1997. The Teleface project - Multimodal speech communication for the hearing impaired. In *Proceedings of Eurospeech'97*. Rhodes, Greece.

[7] Öhman, T. 1998. An audio-visual speech database and automatic measurements of visual speech. *TMH-QPSR, KTH*, 1-2/1998: 61-76.