Animated Lombard speech: Motion capture, facial animation and visual intelligibility of speech produced in adverse conditions

Simon Alexanderson, Jonas Beskow*

KTH Speech, Music and Hearing, 100 44 Stockholm, Sweden

Received 24 October 2012; received in revised form 7 February 2013; accepted 25 February 2013
Available online 5 March 2013

Abstract

In this paper we study the production and perception of speech in diverse conditions for the purposes of accurate, flexible and highly intelligible talking face animation. We recorded audio, video and facial motion capture data of a talker uttering a set of 180 short sentences, under three conditions: normal speech (in quiet), Lombard speech (in noise), and whispering. We then produced an animated 3D avatar with similar shape and appearance as the original talker and used an error minimization procedure to drive the animated version of the talker in a way that matched the original performance as closely as possible. In a perceptual intelligibility study with degraded audio we then compared the animated talker against the real talker and the audio alone, in terms of audio-visual word recognition rate across the three different production conditions. We found that the visual intelligibility of the animated talker was on par with the real talker for the Lombard and whisper conditions. In addition we created two incongruent conditions where normal speech audio was paired with animated Lombard speech or whispering. When compared to the congruent normal speech condition, Lombard animation yields a significant increase in intelligibility, despite the AV-incongruence. In a separate evaluation, we gathered subjective opinions on the different animations, and found that some degree of incongruence was generally accepted.

© 2013 Elsevier Ltd. All rights reserved.

Keywords: Lombard effect; Motion capture; Speech-reading; Lip-reading; Facial animation; Audio-visual intelligibility

1. Introduction

Humans are extremely versatile in their way of compensating for external factors in spoken communication. In the presence of acoustic noise, humans seamlessly adapt their speech production and perception strategies in order to compensate for the acoustic external conditions and maintain communication. According to the theory of Hyper–Hypo articulation (Lindblom, 1990), speakers tend to economize their speech production with the goal to make themselves understood in a particular communicative situation. Speech produced in noise exhibits not only increased loudness, but also larger articulatory movements (Fitzpatrick et al., 2011).

Adaptation occurs not only in the production system but also in speech perception. When the acoustic channel is subject to disturbances, humans tend to rely more on other information sources, such as vision. It is well known that the visual modality has a strong influence on speech perception, and can even alter the perception in its entirety (McGurk and MacDonald, 1976). In acoustically adverse conditions, visible speech enhances speech comprehension.

* Corresponding author. Tel.: +46 8 790 8965.
E-mail addresses: simonal@kth.se (S. Alexanderson), beskow@kth.se, jbeskow@gmail.com (J. Beskow).
This audiovisual speech enhancement effect is particularly important for people with hearing impairments, but is equally present in normal hearing persons in the presence of external noise (Summerfield, 1992).

While both of these effects have been extensively studied in isolation, the combined effect – i.e. how audiovisual speech perception is affected by the presence of noise during speech production, has been considerably less investigated. According to Campbell and Mohammed (2010), speakers speaking using normal voice, as opposed to whispering or shouting, are usually judged easiest to speech-read, and Lombard speech can arguably be considered a form of shouting. On the other hand, Kim et al. (2011) found that the audiovisual speech enhancement in a speech in noise audiovisual sentence intelligibility task, was greater for speech recorded in noisy conditions than for speech recorded in quiet conditions. Similar results were reported by Vatikiotis-Bateson et al. (2007), in a study following up on Vatikiotis-Bateson et al. (2006), who found no support for larger AV enhancement for Lombard speech.

In our work we are concerned with talking characters that can be automatically animated from either text or speech with high enough accuracy that they provide substantial visual intelligibility enhancement. While several such systems have been presented (see Bailly et al. (2003) for an overview), none of these have the capability of modelling the types of articulatory variations associated with Lombard speech. Given the reported increases in AV intelligibility for Lombard speech, we want to investigate whether these enhancements carry over to our talking head animation, when we drive the animation from motion capture. More interestingly, we want to see if animation modelled after Lombard speech can be used to enhance the visual intelligibility in a general sense, i.e. even for a non-Lombard voice, in spite of the potential voice/face mismatch that this would entail.

The ability of a virtual character to adapt articulatory effort can be important for other reasons as well. Recent work in statistically driven speech synthesis (c.f. Raitio et al., 2011) include the capability of adapting the voice quality to Lombard speech. In the interest of coherence, we would expect the same capability to be present in a talking head animation, in the interest of an audio-visualy coherent user experience. Similarly, if the animated face is being driven from real speech, as in Salvi et al. (2009), it would be desirable to have the ability to adapt to the speaking style of the input voice.

In the on-going Lipread project, we are designing a talking face to be used in an e-learning environment for training of speech-reading. In a training situation such as this one, it is important to be able to modify the animation style to give the learner a more diverse training situation. In summary, having the ability to model speaking style in terms of articulatory effort would potentially give us two things: facial animation that can conform to the style of a given real or synthesized voice, and a visual speech synthesis system where the degree of articulatory effort (and potentially the intelligibility), can be explicitly controlled.

In the current study we used speech in quiet, speech in noise and whispering as a way of eliciting different degrees of articulatory effort from our talker while he was recorded using motion capture equipment. Then we mapped the motion capture data onto an animated face rig and produced animated versions of the recorded sentences. Finally we conducted a perceptual evaluation study, where we compared the intelligibility of the generated facial animation against the natural video as well as audio-only versions.

2. Audio, video and motion corpus

We have recorded a multimodal (audio, video, motion capture) corpus covering different speaking styles. The purpose of the corpus is twofold. Firstly, we want to investigate the visual intelligibility of an animated talker with different speaking styles as outlined in Section 1, which is the main focus of this article. To this end we need suitable material for intelligibility testing. Secondly, we want the corpus to function as training data for a statistically based visual speech synthesizer, so we need a phonetically balanced data set. To satisfy these goals we have chosen to record a Swedish sentence set previously developed for the purpose of audiovisual intelligibility testing (see Öhman, 1998) as well as a set of nonsense VCV words.

2.1. Data recording

The speaker was a male Swedish actor who was seated face to face with a listener (approx. 2 m away), and was instructed to read short sentences and words from a monitor, and make sure that the listener understood what was being said. Both listener and speaker wore headphones (Sennheiser HD 600), where they could hear their own speech as picked up by a common omni-directional microphone, at a level that was pre-adjusted to roughly compensate for the
Fig. 1. Screenshots from the recordings. (a) Quiet, (b) noise 80 dB, (c) whispered, and (d) marker data from head, face and sync device.

Table 1
Average inter-lip velocities and durations for all Swedish sentences.

<table>
<thead>
<tr>
<th></th>
<th>Quiet</th>
<th>Noisy</th>
<th>Whispered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed (mm/s)</td>
<td>16.6</td>
<td>34.0</td>
<td>32.0</td>
</tr>
<tr>
<td>Sentence duration (s)</td>
<td>1.56</td>
<td>2.45</td>
<td>2.68</td>
</tr>
</tbody>
</table>

attenuation of the headphones. An optional stationary brown noise signal was also fed to the headphones, at different levels throughout the recording. In order to determine an appropriate noise level for elicitation of Lombard speech, we made trial recordings at 70, 80 and 90 dB SPL. We chose to use 80 dB, since at the 70 dB level, communication was found to be too easy and at the 90 dB level too difficult.

The speaker’s facial movements were recorded by a 10-camera NaturalPoint OptiTrack optical motion capture system operating at 100 frames/s. The speaker was equipped with 37 reflective facial markers + 4 on the head, following the recommended marker placement template from NaturalPoint. In addition, HD-video was captured using a JVC GZ-1 video camcorder. Speech was recorded via a Studio Projects C1 large diaphragm condenser microphone and an RME FireFace 80 external sound card. In order to synchronize the motion capture and the audio, a custom device was constructed, featuring three switchable IR LEDs. When switched on, the LEDs would show up as markers in the motion capture system, at the same time producing an electrical pulse in one input channel of the sound card. The same sync pulses were fed to the external-mic input of the camcorder, thus allowing for precise and fully automated post synchronization of all data streams.

A set of 180 short Swedish sentences was recorded under three different conditions: Quiet (Q), Noisy (N) and Whispered (W). Quiet is the baseline condition, where no noise was presented in the headphones. In the Noisy condition, brown noise at the level of 80 dB SPL was presented in the headphones of both speaker and listener. In the Whispered condition, no noise was presented, but the speaker was instructed to keep his voice to a whisper (no voicing), and still try to make himself understood to the listener. This was done in an attempt to elicit exaggerated lip movements (Fig. 1).

2.2. Data post-processing

After the motion capture was recorded the markers were identified and tracked, generating trajectories for each marker. The captured data did not show any occlusions and there was a small amount of erroneously reconstructed markers (ghost markers). To smooth the data, we used a local regression method using a 2nd degree polynomial model and window size set to 5 frames (0.05 s). We verified by manual inspection that the smoothing did not flatten peaks and valleys in the motion data causing e.g. loss of lip closing. The facial marker motion was then stabilized to the head markers, resulting in separate streams for the global head rotations and translations, and local positions of the face markers expressed in a reference frame fixed to the head.

As a final step, the data was cut into segments based on the sync-signal injected by the switched LEDs. A first analysis was made by studying the distance between upper and lower lip. Fig. 2 shows this distance for one Swedish sentence, for the quiet and noisy (80 dB) conditions. As expected, the lip movements exhibit much larger amplitude in the noisy conditions than in the quiet. Table 1 shows that the average inter-lip velocity is lowest in the quiet condition and increases with noise level. The whispered condition exhibits almost twice the velocity as the in the quiet case. Another important difference is the reduced speaking rate in the noisy and whispered conditions (see Table 1).
<table>
<thead>
<tr>
<th>Speed (mm/s)</th>
<th>Quiet</th>
<th>Noisy</th>
<th>Whispered</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.6</td>
<td>34.0</td>
<td>32.0</td>
<td></td>
</tr>
</tbody>
</table>

| Sentence duration (s) | 1.56 | 2.45 | 2.68 |

Fig. 2. Inter-lip distance for the Swedish sentence ‘Dom flyttade möblerna’ in three recording conditions: quiet (top), 80 dB noise (middle) and whispered (bottom).

3. Animation

The field of 3D-animation offers a wide range of tools and techniques to animate talking heads, see Deng and Noh (2007) for an overview. The main considerations in our work were choice of 3D geometry, facial parameterization and rendering style. As we plan to use the talking head in real-time applications we chose to adopt standard techniques used in the video-games industry. The facial rig was a 3D model generated from a commercial system, FaceGen (Singular Inversions Inc.), having a blendshape (morph target) based facial parameterization, and the rendering was done in
Open scene graph (OSG). To produce the video clips for our study, the images from OSG (640 × 480 resolution) were exported frame-by-frame and compressed with a MPEG-4 codec to Quicktime format.

3.1. 3D representation of head and face

FaceGen generates 3D models of human heads by modification of various parameters for sex, age, race etc. For our purposes, one of the key features of FaceGen was the support for generation of 3D heads fitted to photographs. A 3D model closely matching the physiognomy of the captured speaker simplifies the process of cross-mapping motion data to target animation, and reduces the effects of appearance in studies where the animated head is compared to natural video.

We used one frontal image and two profile images of the captured speaker to generate a 3D model and colour texture map, see Fig. 3. The standard FaceGen models have a blendshape based facial parameterization consisting of 41 blendshapes. Our first tests showed that the facial rig was unable to reproduce subtler expressions and asymmetric movements present in the motion capture data. The main problem was that the blendshapes were modelled to influence the whole mouth region globally, lacking a separate control of left, right upper and lower regions. To gain a more fine-grained control of the facial animation some of the original blendshapes influencing the mouth region were split into left/right upper lip/lower lip, using a smooth transition based on a cosine function. The splitting procedure resulted in a total of 62 blendshapes, which were used in the animation mapping process.

3.2. Transfer of performance data to animation

Mapping captured data to 3D facial animation is a nontrivial task that has been the focus of much research, see Pighin and Lewis (2006). Despite recent improvements in technology and research, most high quality animations seen in movies or computer games still require extensive manual post-processing by trained animators. Inspired by Choe (2001), we developed a mapping procedure that requires little manual intervention. The procedure consists of a basic algorithm adjusting the facial rig to fit the marker locations in each frame, see Section 3.2.1, and an extension to enforce bilabial- and labiodental occlusions, see Section 3.2.2.

3.2.1. Mapping

To obtain a one-to-one mapping of feature points of the captured data and the target model, a set of landmark vertices, or virtual markers, were manually picked from the target face model at the corresponding locations of the captured markers. Next, we performed a spatial alignment of the captured data with the target model by applying a rotation, scale and translation transformation. The transformation was calculated at a reference frame, where the speaker’s face was in neutral pose similar to the base shape of the target model, by minimizing the distance errors from the marker positions to the corresponding virtual markers of the target face.

After the motion data was aligned to the model, we looked for the optimal blendshape parameters for the virtual markers to follow the trajectories of the captured markers. At each frame, we calculated the marker dislocations from
the reference neutral pose and used an iterative algorithm to find the 62-dimentional vector of blendshape weights minimizing the total distance error to the corresponding virtual marker dislocations

$$e^2 \sum_j |v^*_j - v_j|$$

where $v^*_j$ and $v_j$ are the dislocation vectors for virtual marker and captured marker $j$ respectively.

3.2.2. Articulatory target constraints

For phoneme production, some phonemes have more visually critical constraints than others. In order to provide good intelligibility, these constraints should be met by the animations. Typically bilabial phonemes enforce upper lip-to-lower lip constraints and labiodental phonemes lip-to-upper-teeth constraints, while vowels have less visibly salient constraints.

To improve the basic mapping algorithm described above we added weights to the error Eq. (1) giving a larger penalty for errors from the upper and lower lips at the presence of a bilabial- or a labiodental occlusions. In order to reliably detect such occlusions from the marker data, we used a probabilistic approach informed by phonetic knowledge.

We used an automatic HMM-based forced alignment system (Sjölander, 2003) to derive time-aligned phonetic transcriptions. Tree Gaussian mixture models (GMMs) were trained to predict vowels and bilabial- and labiodental phonemes using 12 spatial features (the local coordinates $x$, $y$, $z$ of the markers on the upper lip, lower lip and mouth corners) from the motion capture data. The probabilities from the GMMs were applied to our minimization algorithm as outlined above. A detailed description on the procedure will be presented in forthcoming work.

The extended mapping algorithm was applied to all recorded sentences and the animations were rendered in OSG. Images from the final videos are shown in Fig. 4.

3.2.3. Time alignment of articulatory parameters

In order to make a direct comparison between different articulatory conditions for the same audio file, we constructed a time alignment between the different renditions of the same sentence (Q–N and Q–W). We used phonetic label files from forced alignment (see above) and matched up the labels in each sentence pair using longest common subsequence (LCS) matching. This established a number of temporal correspondence points between the two label files (typically before and after every phone that the two label files have in common). Next, linear interpolation is used to map every time point at 10 ms intervals in the Q sentence to a corresponding time point in the N or W sentence. These time points can later be used during animation to find the correct position in the Lombard sequence, given a position in the Q sentence.
4. Method

4.1. Experiment 1: Audiovisual intelligibility test

In order to assess the visual enhancement effect of the different renditions of the talking head and compare these to the natural recorded video of the speaker, we conducted an audiovisual sentence intelligibility experiment. We used acoustically degraded sentences, where the content of the acoustic sentence is partially intelligible when using audio signal only. This was done in order to avoid ceiling effects. The sentences were presented in different audio only or audiovisual conditions, in total 11 conditions were used (see below). The intelligibility, measured as the number of correctly recognized words that the subjects report is compared across conditions.

In the experiment, the audiovisual stimuli consisted of a collection of short Swedish sentences, which vary in length between three to six words, with a basic everyday content. e.g., “Den gamla räven var slug” (The old fox was cunning).

The audio signal was processed using a vocoder excited by white noise (Shannon et al., 1995) to reduce intelligibility. This vocoder applies band-pass filtering and replaces the spectral details in the specified frequency ranges with noise. The vocoder has been previously applied in similar experiments exploring the effects of audio-visual intelligibility (Engwall & Wik, 2009; Al Moubayed et al., 2010).

In order to avoid any floor or ceiling recognition rate effects, we wanted to ensure an intelligibility rate between 25% and 75%. A pilot test was run to determine the number of frequency bands to use in the vocoder (higher number of frequency bands leads to a more intelligible speech signal). For speech in quiet and speech in noise, three frequency bands were used, and for whispered speech six bands were used.

The test included a total of eleven presentation conditions: each of the three recording conditions Q, N and W were presented in audio only, animated video, and natural video conditions, yielding nine conditions. In addition to this, two incongruent animated video conditions were added: speech in quiet with time-warped articulation from speech in noise, and speech in quiet with time-warped articulation from whisper. This was done in order to facilitate a direct comparison of the visual enhancement of the different articulation styles over the same audio only condition. The conditions are presented in Table 2.

Eleven subjects ranging in age from 18 to 38 years took part in the test. All subjects were native speakers of Swedish and reported normal hearing and normal or corrected to normal eye sight. The stimuli were presented in 11 blocks of 15 sentences per group and every block was only used for one condition. Sentence blocks and presentation conditions were counter balanced between the subjects, in such a way that none of the 11 subjects received the same combination of sentences and presentation conditions. This was done in order to avoid interaction effects between the sentence difficulty and the presentation condition. Presentation order for the 11 different conditions was also rotated between the subjects so that no subject started with the same presentation condition.

Prior to the experiment, subjects were enrolled in a training session, to adjust to the testing situation and to the noise excited vocoder audio quality. This was done in order to avoid any training effects during the actual experiment. During training, subjects were allowed to listen to the degraded audio file as many times as they wished, and feedback was given to them with the correct content of the audio sentence. Typically during training, subject’s performance was

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>audioQ</td>
<td>Speech in quiet</td>
<td>Auditory-only</td>
</tr>
<tr>
<td>audioN</td>
<td>Speech in noise</td>
<td>Normal AV (audio + natural video)</td>
</tr>
<tr>
<td>audioW</td>
<td>Whisper</td>
<td>Congruent animated AV (audio + animated video)</td>
</tr>
<tr>
<td>videoQ</td>
<td>Speech in quiet</td>
<td>Incongruent animated AV (audio + time-warped animated video)</td>
</tr>
<tr>
<td>videoN</td>
<td>Speech in noise</td>
<td></td>
</tr>
<tr>
<td>videoW</td>
<td>Whisper</td>
<td></td>
</tr>
<tr>
<td>animQ</td>
<td>Speech in quiet</td>
<td></td>
</tr>
<tr>
<td>animN</td>
<td>Speech in noise</td>
<td></td>
</tr>
<tr>
<td>animW</td>
<td>Whisper</td>
<td></td>
</tr>
<tr>
<td>audioQ-animN</td>
<td>Audio: Speech in quiet, animation: speech in noise</td>
<td></td>
</tr>
<tr>
<td>audioQ-animW</td>
<td>Audio: speech in Quiet, animation: whisper</td>
<td></td>
</tr>
</tbody>
</table>
initially very poor, often close to 0% accuracy, but performance increased quickly during training, converging to about 40–60% at the end of the training session. The average length of the training sessions was nine sentences. During the experiment, subjects only had the chance to listen to the stimulus once, and were never provided with the correct answers. The sentences used during the training session were not used in the experiment.

Altering articulation style may not only affect intelligibility, but also the willingness to look at or interact with a talking head. To evaluate the subjective reception of the different articulation styles we performed a second experiment. For this experiment we prepared five video clips of the same sentence, one for each articulation style (animQ, animN, animW, audioQ–animN and audioQ–animW) from experiment 1. Unlike experiment 1, we did not use a vocoded speech signal, but the original sound from the recorded session normalized to a comfortable listening volume. This sound did not include the noise presented to the speaker during the recording.

The five video clips were presented to 17 subjects who were asked to fill in a questionnaire and to rate the animations in three different aspects: how well the motion matches the sound (coherence), if the animations are disturbing or easy to take in (acceptance), and to what extent the animations are perceived as eerie (eeriness), see Table 3. The rating where given in a scale from 1 to 5, where 1 is the lowest score, and 5 the highest. First, the subjects were instructed to read all the questions, then they were presented with all video clips twice in a consecutive order, and finally they viewed the video clips again one by one while filling in the questionnaire. The presentation order of the videos was rotated for each subject.

5. Results

5.1. Results from experiment 1

The audio-visual intelligibility scores of each sentence were gathered for all subjects. Each of the sentences contained three key content words, generating a scale from 0 to 3 for the number of correct words picked up by the subject. The mean scores for the presentation modes are presented in Fig. 5. To investigate the significance rates of the experiment,
the results were analysed in a univariate ANOVA with a LSD post hoc test. In Table 4 we highlight some of the results from the post hoc test.

Effects of recording condition: as can be seen in Fig. 5 and Table 4, the speech recorded in noise gave a significantly higher intelligibility than the speech recorded in quiet condition in all presentation groups (auditory-only, normal AV and congruent animated AV). Note that the whispered condition used a different vocoder setting (see Section 4.1) and should not be compared with the other recording conditions in each group.

Effects of visual stimuli: The results show a significant increase in intelligibility when adding a visual modality to the speech. This is seen over all recording conditions (Q, N and W), the effect being greatest comparing auditory-only to normal AV, but also when comparing auditory-only to animated AV, where animN and animW showed insignificant changes to the corresponding natural videos in videoN and videoW.

A notable result is the intelligibility scores for the incongruent cases audioQ–animN and audioQ–animW (normal speech with time-aligned Lombard or whispered articulation). Both these conditions show a higher intelligibility rate than the animQ condition.

5.2. Results from experiment 2

The results from experiment 2 were analysed in a univariate ANOVA with a LSD post hoc test. The mean scores for the three questions are presented in Fig. 6. As can be seen in the figure, the coherence ratings for the congruent animated AV conditions (animQ, animN, animW) were highest, while the coherence rating for the audioQ–animW condition was significantly lower than all the others ($p < 0.05$). Also the acceptance-rating was significantly lower for

![Fig. 6. Average ratings (1–5) of coherence, acceptance and eeriness for the presented videos. The error-bars are set to one standard deviation.](image)
the audioQ–animW condition than the others ($p < 0.05$). In the eeriness-ratings audioQ–animW was perceived as most eerie followed by the whispered animation (animW).

6. Discussion

6.1. Experiment 1

As expected, the results from experiment 1 show that Lomard speech is more resilient to auditory disturbances than normal speech. This applies to both the auditory-only and audiovisual domain. While it is not possible to make direct comparisons between whispered speech and the other speech conditions (quiet and in-noise) due to different number of vocoder bands used, it is found that visual information complementing speech gives an increase of intelligibility in all three recorded conditions (speech in quiet, speech in noise and whispered speech). By comparing the results from the natural videos and the animated talking head, we found that the enhancements from the visual cues are well translated in the speech in noise and the whispered speech conditions, but less satisfactory in the speech in quiet case. An interesting finding for the incongruent animated AV conditions (audioQ–animN and audioQ–animW), is that increased visual articulation combined with normal speech is perceived as more intelligible than normal speech and articulation (animQ). This result can be exploited e.g. for the purpose of building talking heads with independently varied auditory and visual style according to desired application, which is one of the goals of this work as outlined in the introduction. For lip-reading training applications, only the visual modality may be varied to modify the difficulty level of the lip-reading task. And for an embodied agent set in noisy environment, both auditory and visual style could be tailored to maximize intelligibility for the current noise level.

6.2. Experiment 2

The most interesting findings from experiment 2 were made comparing the animations with incongruent source of audio and visual data (audioQ–animN and audioQ–animW) to the animations where audio and motion originates from the same recording.

The ratings show that the animation with normal voice combined with whispered articulation (audioQ–animW) was in average considered worst in matching of motion to speech, least easy to take in, and most eerie. This was not surprising since this condition showed most artefacts arising from the time-warping process; the whispered recordings showed most extensive articulation and longest duration times, and when the animations were aligned to normal speech rate the time warping generated some unnaturally fast lip movements. But the animation with normal voice and speech in noise articulation (audioQ–animN), also having the same artefacts but to a lesser degree (as seen in the coherence-rating), was considered equally easy to take in as the congruent conditions, and was considered less eerie than both the congruent whispered- and speech in noise animations. This suggests that there is a certain tolerance to mismatch of speech and motion without negative effects on the perception.

The different conditions having different vocal styles may have affected the ratings. In the quiet condition the voice of the speaker was very calm and relaxed, while the shouting voice from the noisy condition and the whispered voice may evoke different emotional responses. As a consequence the speech in noise animation may have been biased towards more disturbing in question 2. Likewise, the whispered voice in the animW condition may be culturally seen as more eerie, which may be the cause of the relatively high eeriness rating of this animation.

7. Conclusions

In this paper we have investigated methods to elicit speaking styles with varied degrees of articulatory effort, with the primary goal of collecting motion data for highly accurate and intelligible talking face animation. In order to ensure that the visual intelligibility gain transfers from the recorded speaker to the animation, we conducted an audio-visual intelligibility test where we compared the video recorded talker and the animated talker, for the different speaking styles (speech in quiet, speech in noise and whisper). We can conclude that as expected, all visual conditions led to better intelligibility than the acoustic counterparts alone. Equally expected, we found that the speech in noise gave higher intelligibility than the speech in quiet condition, which is in line with previous findings. This effect is due to
the increased articulatory gestures as well as a slower speaking rate and more the elaborate phrasing that could be observed for our talker in the speech in noise and especially in the whisper condition.

More importantly, we found that the visual intelligibility of our recorded talker transferred well to the animated talker, as there was no significant difference in intelligibility between these two for the speech in noise and whisper conditions. This implies that the animation mapping procedure described in Section 3 is satisfactory, at least for these two types of speech. For the speech in quiet condition, the natural talker was more intelligible than the animated one. A possible explanation for this is that speech in quiet exhibits much smaller articulatory movements than speech in noise, and thus relies on subtler cues, which may not transfer as well to the animation as the larger movements. Still, there is a considerable benefit from seeing the speech-in-quiet animation over the audio alone case.

Another important finding is that for speech in quiet, there is indeed a significant visual intelligibility gain to be had by driving the animated talker with the speech in noise or whisper data, but time aligned to speech in quiet audio. What this means is that the benefit resulting from larger and more elaborate articulatory movements can to some extent be separated from the intelligibility boost coming from the phrasing and durational changes introduced in the more exaggerated speaking styles. It also means that although we introduce incongruence between audio and animation in this case (i.e. they are not taken from the same articulatory condition) subjects still get the added benefit of these lip movements. This is also what we could expect based on previous experience with animated talking faces – even if the animated movements do not originate from the same speaker as the voice, which most of the time is case for speech- or text driven facial animation, there can still be a considerable intelligibility gain from such animated faces (Salvi et al., 2009). Still, the subjective evaluation study seems to suggest that there is a limit to the amount of temporal distortions that can be performed to fit one animation to a certain audio clip before it is perceived as unnatural. The speech in noise animation is accepted together with the speech in quiet audio, whereas the whisper animation, which differs more from the quiet animation in terms of articulatory dynamics, suffers more from time compression and is not received well in the subjective test. In other words, there seem to be some tolerance for mismatch in sound and motion in the perception of an animated talking head, but too grave temporal distortions (e.g. unnaturally fast movements) is harder to accept.

The next step in this line of work will be to build a statistically based visual speech synthesizer trained on the corpus described in this paper. In this system we will try to implement a continuous control over articulatory effort that can be used to increase/decrease visual intelligibility (useful in the speech-reading training) or to adapt to adverse acoustic conditions, ideally in combination with a text to speech system with corresponding abilities. This is likely to be a useful feature for e.g. multimodal dialogue systems or social robots operating in acoustically dynamic or challenging conditions.

Acknowledgements

This work was been carried out with the support of the Lifelong Learning Programme of the European Union, project 518970-LLP-2011-NL-KA2-KA2MP (LipRead) and the Swedish Research Council, project VR 2010-4646 (Massive).

References


Campbell, R., Mohammed, T.J.E., 2010. Speechreading for information gathering: a survey of scientific sources. Deafness Cognition and Language (DCAL) Research Centre Division of Psychology and Language Sciences, University College London.


