J Multimodal User Interfaces (2010) 3: 299–309
DOI 10.1007/s12193-010-0054-0

ORIGINAL PAPER

Auditory visual prominence

From intelligibility to behavior

Samer Al Moubayed · Jonas Beskow · Björn Granström

Received: 16 November 2009 / Accepted: 14 September 2010 / Published online: 30 September 2010

Abstract Auditory prominence is defined as when an acoustic segment is made salient in its context. Prominence is one of the prosodic functions that has been shown to be strongly correlated with facial movements. In this work, we investigate the effects of facial prominence cues, in terms of gestures, when synthesized on animated talking heads. In the first study, a speech intelligibility experiment is conducted, speech quality is acoustically degraded and the fundamental frequency is removed from the signal, then the speech is presented to 12 subjects through a lip synchronized talking head carrying head-nods and eyebrows raise gestures, which are synchronized with the auditory prominence. The experiment shows that presenting prominence as facial gestures significantly increases speech intelligibility compared to when these gestures are randomly added to speech. We also present a follow-up study examining the perception of the behavior of the talking heads when gestures are added over pitch accents. Using eye-gaze tracking technology and questionnaires on 10 moderately hearing impaired subjects, the results of the gaze data show that users look at the face in a similar fashion to when they look at a natural face when gestures are coupled with pitch accents opposed to when the face carries no gestures. From the questionnaires, the results also show that these gestures significantly increase the naturalness and the understanding of the talking head.

S. Al Moubayed (✉) · J. Beskow · B. Granström
Center for Speech Technology, KTH, Lindstedtsvägen 24,
10044 Stockholm, Sweden
e-mail: sameram@kth.se

J. Beskow
e-mail: beskow@kth.se

B. Granström
e-mail: bjorn@speech.kth.se

Keywords Prominence · Visual prosody · Gesture · ECA · Eye gaze · Head nod, eyebrows

1 Introduction

There is currently considerable interest in developing animated agents to exploit the inherently multimodal nature of speech communication. As animation becomes more sophisticated in terms of visual realism, the demand for naturalness in speech and gesture coordination increases.

It has long been recognized that visual speech information is important for speech perception [1, 2]. There has been an increasing interest in the verbal and non-verbal interaction between the visual and the acoustic modalities from production and perception perspectives. Studies have reported possible correlations between acoustic prosody and certain facial movements. In [3], correlation between f0 and eyebrow movements is discussed. In [4], correlations between f0 movements and head-movement dimensions are reported and such movements are found to increase speech-in-noise intelligibility. Such coupling of movements in the acoustic and the visual modalities is usually highly variable, but an understanding of the redundancy of information in these two modalities can greatly help in developing audio-visual human-human and human-machine interfaces to guarantee a maximum amount of interaction. More recently, the contribution of facial prosody to the perception of the auditory signal has been investigated. In [4] it was shown that head movements can be used to enhance speech perception by providing information about the acoustic prosodic counterpart; in [5] and more recently in [6], it is shown that even movements of the top of the head can aid speech comprehension. Moreover, people can highly discriminate the acoustic prosody of an utterance only by looking at a video showing the top of the head. All these studies suggest a highly shared production and perception of prosody between the acoustic and the facial movements. However, knowledge on how to quantify this strong relation, and how to deploy it in systems is still highly understudied.

One of the prosodic phenomena which attracts much focus is prominence. Prominence is defined as when a linguistic segment is made salient in its context [7]. Words (or longer or shorter segments) can be made prominent to convey information such as contrast, focus [8], and information status [9]. Hence, the communication of prominence can impact upon the interpretation of a word or phrase, and affect the speech comprehension.

Recent studies have focused on the relation between the visual modality (the face) and acoustic prominence [10]. In [11], results on Swedish showed that in all expressive production modes, words which are produced with a focal accent exhibit greater variation in the facial parameters movements (articulators, eyebrows, head, etc.) than when the word is in a non-focused position. In [12], visualizing eyebrow movements and head nods on a talking head is found to be a powerful cue to enforce the perception of prominence. In [13, 14], an investigation on the interaction between the acoustic and the visual cues of prominence is conducted. The results of this study suggest that, during production, when a word is produced with a visual gesture, the word is also produced with a higher acoustic emphasis. From a perception perspective, the results suggest that when people see a visual gesture over a word, the acoustic perception of the word's prominence is increased. In [15], it is shown that visual movements do not only enhance the perception of focus, but can even decrease the reaction time in detecting it; and in [16] it was shown that focus can be detected not only through head or eyebrow movements but also through lip movements.

From this various evidence on the interaction between facial movements and acoustic prominence, a natural question is whether this correlation between the acoustic and the visual prominence can extend beyond the role of triggering the perception of prominence visually as a prosodic unit, that is if it might play a role in speech comprehension by providing information about the linguistic structure underlying the segments which receive acoustic and visual prominence. Several studies have shown that perceiving facial movements beyond the articulatory ones can increase speech intelligibility, as in viewing head movements in [4]. These studies however, do not give information on which particular movements provide these effects.

In the first experiment in this paper we investigate the effects of visual cues to prominence and pitch accents on speech intelligibility. In the second experiment, we study the non-verbal interaction effects of these cues when synthesized on talking heads and synchronized with pitch accents, by studying the eye gaze behavior of hearing impaired subjects.

The paper is organized as follows, in Sect. 2 we briefly introduce the prominence prosodic model in Swedish. In Sect. 3, a study on the effects of visual cues to prominence is conducted by presenting acoustically degraded sentences to 12 subjects, and the intelligibility of these sentences is measured under different visual conditions. In Sect. 4, we present a study on the effects of synthesizing gestures over pitch accents on the behavior of the eye gaze of 10 hearing impaired subjects, and results from questionnaires are presented. In Sect. 5, we discuss various aspects and implications of these studies, and in Sect. 6, we conclude the paper.

2 Acoustic prominence in Swedish

Prominence is defined in [7] as when a segment (a syllable or a word) stands out of its context. Others use it as the perceptual salience of a linguistic unit [17]. Prominence as a non-verbal prosodic category is language independent, nevertheless, its realization through acoustic parameters is language dependent. Since the studies conducted in this work are in Swedish, we present an overview of the acoustic correlates to prominence in the Swedish language. In Swedish, prominence is often categorized with three terms corresponding to increasing levels of prominence: *stressed, accented* and *focused*. Research has reported that the most consistent acoustic correlate of stress in Swedish is segmental duration [18]. In addition, overall intensity differences have also been studied among the correlates of stress, although these differences may not be as consistent as the durational differences [18]. As to *accented* syllables, according to the Swedish intonation model in [19], the most apparent acoustic difference between an accented and an unaccented foot is the presence of an f0 fall, referred to as a word accent fall. Thus, accent as a higher prominence level than just stress is signaled mainly by f0, although an accented foot is usually also longer than an unaccented one [20]. Finally, in focal accent, which is the highest level of prominence, the primary acoustic correlates for distinguishing 'focused' from 'accented' words is a tonal one—a focal accent or a sentence accent rise following the word accent fall [19]. However, this f0 movement is usually accompanied by an increased duration of the word in focus, [21], and by moderate increases in overall intensity [22]. In our studies, we deal with prominence as a perceptual prosodic phenomenon on a continuous level, which is separate from its underlying linguistic context. Looking at prominence from this stand point has been suggested previously in [23].

The experiments on the visual effects of prominence, presented here, are applied to the Swedish language. Although the Swedish prominence model is structurally different than that of other languages, this difference does not necessarily lead to language dependent perception of the visual cues of prominence. Previous research has tried to investigate the functional differences in accentuation using eyebrows across Dutch and Italian [24]. Differences are indeed found, but it is argued that these differences can be reduced to prosodic differences between these languages, but not to language dependent perception of the visual movements. However, it is interesting to further study whether these effects are consistent across languages.

3 Experiment 1: visual prominence and speech comprehension

As mentioned earlier, several studies support a strong relation between the auditory and the visual modalities in perceiving prominence. These studies though, only report that the perception of prominence as a non-verbal prosodic phenomenon is highly aided by certain visual movements. Hence, the visual modality supports the realization of the acoustic prominence which exists in the audio stream.

Since prominence is manifested differently depending on the linguistic segment underlying it, and perceiving prominence aids speech comprehension, the question asked in this study is whether perceiving cues of prominence through the visual modality can increase speech intelligibility when the acoustic prominence is degraded. And hence, visualizing prominence on the talking head will not only be useful to transmit/correlate its acoustic counterpart, but also support speech perception of the acoustic signal.

To investigate such effects, a speech intelligibility experiment is conducted with the help of a lip synchronized animated talking head.

3.1 Method and setup

Computer synthesized talking heads have been progressively developing, offering the possibilities for many experimental designs which are not possible otherwise. That is by manipulating and changing the stimuli in one modality while keeping the other modality intact, creating a stimuli setup which explores the effects of specific variables, for example, by manipulating the required variable to measure, and keeping the others static [25]. This is the main reason for the use of a lip-synchronized talking head as the medium for the visual modality in this work. The talking head used has been shown to provide increased intelligibility in several previous studies [26, 27].

This experiment design deploys the approach of presenting human subjects with a degraded speech signal, while they are looking at a talking head [28]. Different sets of sentences receive facial gestures at different timings along the speech signal, and the differences in cross-subject speech intelligibility is studied.

40 semantically complete sentences, ranging in length between 6 and 10 words, were selected from a corpus containing news texts and literature, read by a professional Swedish actor. The corpus contains high-quality studio recordings for the purpose of voice creation for speech synthesis systems.

The speech files of the 40 sentences were force-aligned using an HMM aligner [29] to guide the talking head lip movement generation procedure [30]. The audio was processed using a 4-channel noise excited vocoder [31] to reduce intelligibility. The number of channels was decided after a pilot test to ensure an intelligibility rate of between 25% and 75%, that is to avoid any floor or ceiling limit effects.

All the sentences were presented to subjects with an accompanying talking head. The first 10 sentences were presented without any facial gestures, as a training session, to eliminate any quick learning effect for the type of signal degradation used. The 30 sentences remaining were divided into 6 groups; every group contained 5 sentences, with a balanced number of words in each group (35–40 words). For each group, 6 different visual stimuli were generated (details in Sect. 2.3). These groups were systematically permutated among 12 normal hearing subjects with a normal or corrected to normal vision, so that every subject listened to all 30 sentences, but with each group containing different visual stimuli. During the experiment, the sentences were randomized for every subject. The subjects had one chance to listen to the sentence (while looking at the talking head), and then type in a text field what they understood from the signal.

3.2 Marking prominence

As described in Sect. 2, focal accent is the highest level of prominence. According to the Swedish prominence model, the acoustic correlates to focal-accent can be distributed over more than one syllable in a word. This is done by having a word accent fall, followed later by an accent rise. This is more evident in poly-syllabic words (e.g. compound words). In addition to that, the acoustic correlates, mainly realized as increased syllabic and word duration and f0 movements, can be extended to affect the whole word under focus [32, 33].

For the purpose of this study, the gestures to be included in the stimuli are fixed in length and amplitude and since visual correlates to prominence must be synchronized with their acoustic counterpart (for a study on the effects of timing and shift of prominence gestures see [12]), we decided to limit the size of the focused segment from the whole focused word to its most prominent syllable, and hence, gestures are associated with prominent syllables rather than words.

To establish that, a native Swedish speech expert listened to all the 30 test sentences, and marked them temporally with prominence. In this study, the annotations of only one annotator were token, which reflects a specific subjective perception of prominence (a study in [34] discusses the agreement between annotators on prominence on different levels using the same data-set). By investigating the prominence markers, all sentences received between 1 to 3 prominence marks, and the overall number of marks in the 30 sentences summed to 60.

3.3 Conditions

Following is a detailed description of five of these variants (the sixth condition was a special purpose variant and is left out of this analysis).

It is important to mention here that whenever a sentence received facial gestures, the number of the gestures added to the sentence was always the same in all the visual variants. This is motivated by the fact that, except for the control set which did not receive any gestures, the non-verbal information provided to the signal (deployed here as facial gestures) should be equal among all the different variants of the sentence, and the only variants would be the timing and the type of the gesture.

3.3.1 No gesture (N)

The first condition is 'articulators-only' where the face is static, except for the lips-jaw area for the purpose of phonetic articulation. Every subject had to recognize speech by having one group of sentences in this condition. This condition acts as a control for the rest of the conditions. Figure 1a displays the talking head in the neutral position.

3.3.2 Prominence with head-nods (PH)

In this condition, a head-nod was synthesized in synchrony with the place of the prominence markers in each sentence. The design of the head-nod gesture was near-arbitrary, consisting of subtle lowering and rising to the original location, the complete motion length was set to 350 ms, which is an estimate of the average length of a stressed syllable in Swedish. Figure 1b shows the talking head at the lower turning point of the head nod. Figure 2a displays the vertical trajectory of the head used in the head-nod gesture.

3.3.3 Prominence with eyebrow raise (PEB)

The stimulus in this condition matches the one of the head-nod, except that the gesture in this stimulus is an eye-brow raise, with a matching design in length of trajectories as the head-nod gesture. Figure 1c shows the eye-brow gesture at its top turning point. Figure 2b displays the vertical movement of the eyebrows used in the eyebrows raise gesture.

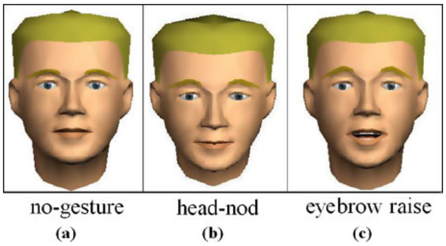

Fig. 1 Snapshots of the talking head in different gestural positions. (**a**) Neutral parameters. (**b**) Peak of the head-nod gesture. (**c**) Peak of the eyebrow raise gesture

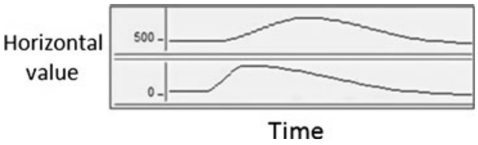

Fig. 2 Plot of the design of the head-nod and eye-brows raise gestures over time. *Top*: the vertical movement of the head. *Bottom*: the vertical movement of the eye-brows

3.3.4 Pitch accent (PA)

A perception experiment in Dutch [35], found that the perception of prominence level is boosted if a pitch accent is accompanied with an eyebrow movement, while it is lowered if the movements are placed on a neighboring word. Studies have reported correlations between f0 and eyebrows movements. Moreover, f0 movements, presented as pitch accents, play a major part in the realization of prominence in Swedish.

In this condition, eyebrow movements were temporally placed in synchrony with steep pitch movements. Each speech file was processed using a sliding window of 150 ms width, with a shift of 10 ms. For f0 tracking, the YIN real-time algorithm is used [36].

The absolute value of the mean delta log f0 was calculated along the f0 signal. According to how many prominence markers each sentence contained, an equal number of markers are placed at the highest peaks of this pitch parameter with a minimum time interval of 350 ms, to avoid overlaps in the gestures (although this constraint was never faced in the sentences). Only pitch movements which were realized over vowels (syllable centers) were coupled with an eyebrow raise in the face. The advantage of this condition is that it is automatically detected, and so it can easily be used for real-time visual speech synchrony in talking agents.

3.3.5 Random eyebrow raise (R)

It is still unclear if a misplacement of a gesture on a non-prominent segment can hinder the comprehension of the speech signal. As noted by previous studies explored above [12, 24], misplacement of prominence movements hinders the perception of prominence on neighboring prominent segments. Nevertheless, the use of gestures might still provide (or confuse) information about the segmental structure of the underlying signal (i.e. word or syllable boundaries). To examine this, eye-brows raise gestures are added randomly on non-prominent syllables with an interval of at least 350 ms to avoid gesture overlap.

3.4 Analysis and results

A string based percentage correct word recognition scoring was applied to the subjects' responses. Strings were scored manually on the word level, which included all function and content words, and the average recognition rate per sentence was then calculated. As a result, every condition received 60 samples (12 subject * 5 sentences per condition).

An ANOVA with normalized recognition rates (Z-scored over subjects) as a dependent variable, and Condition (5 levels: No Gesture, Random, Pitch Accent, Prominence Head-nod, and Prominence EyeBrow raise) as an independent variable was applied. Subjects were not included as a factor since per-subject z-score normalization was applied to effectively remove within-subject differences.

Condition (Treatment) gave a significant main effect: $F(4, 295) = 4.4$, $p < .01$. An LSD post hoc test showed that the conditions: No-Gesture against Prominence Head-nods, No-Gesture against Prominence EyeBrow raise, No-Gesture against Pitch Accents, and Prominence Head-nods against Random eyebrows, are all different using a significance level of 0.05. Table 1 presents the p-value for each pair of conditions. Figure 3 shows the box-plot of the samples per condition against the normalized recognition rates. It is interesting to notice that the only condition which had a significant difference from the random condition (R) was the head-nod (PH) condition, which, in terms of mean recognition rate, had the highest value among all the other conditions (Fig. 3). One possible explanation could be that head-nods actually aid the perception of prominence more than eyebrows, and that more test subjects should be considered to reach a significant difference.

It is important to note that the analysis and results do not show that adding more gestures results in higher intelligibility per sentence or that sentences with more gestures are more intelligible, since sentences were distributed among conditions in a counter balanced scheme. Hence every sentence (with the same number of gestures) was presented in each of the conditions to two different subjects.

Table 1 p-value for every two conditions resulting from the multiple analysis test

Condition	p-value
N*PH	$p < 0.001$
N*PEB	$p < 0.005$
N*PA	$p < 0.05$
PH*R	$p < 0.02$
All other combinations	$p > 0.1$

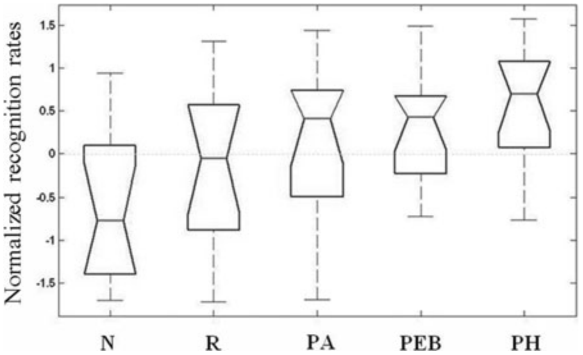

Fig. 3 Box plot of the within-subject normalized % correct word recognition over conditions

3.5 Discussion

These results mainly indicate that adding head-nods or eyebrow on prominent syllables or accented syllables increase speech intelligibility compared to looking at the talking head without gestures at all. On the other hand, adding these gestures randomly over speech, in the same frequency they were added on the prominent syllables, does not increase intelligibility compared to not adding any gestures.

Previously, head-nods have been shown to be a stronger cue in the perception of prominence than eyebrows [12], which seems to be in line with the results in this experiment. Head nods might perhaps be a stronger indication of prominence perceptually due to their larger area in surface motion, and hence require less cognitive effort to realize compared to the perception of eye-brows movements which are realized locally, separately from the lip movements. However, it is hard to investigate from this experiment, in what way the visual realization of prominence has aided the speech perception task.

It is important to stress that this study does *not* claim that these movements (head-nods and eyebrow gestures) are in any way optimal or identical to movements employed by humans to communicate prominence through head and eyebrows (since they are fixed in length, structure, and amplitude), but it is still plausible to assume that these movements to some degree carry the same communicative information contained in human gestures. It also does not claim that, for example, people always provide redundant correlates to acoustic prominence through their head movements and/or eyebrow movements, but it shows that these cues can be of help to speech intelligibility when the acoustic signal is degraded.

Gestures and facial movements are a characteristic of audio-visual interaction, and they might have other roles which affect intelligibility and comprehension than indicating prominence. This could be the reason why there was no significant difference between the random condition and the prominence and pitch accent eyebrow conditions. These facial movements correlate with acoustic prominence, and they have been shown before to play a functional role in face-to-face interaction, such as transmitting or indicating prominence. In addition, as concluded in the previous tests, they support speech intelligibility by presenting prosodic prominence visually. Thus a natural question is: how will these movements support the interaction between humans and talking heads and in what ways would they affect it, e.g. would presenting them increase the naturalness of talking heads?

To study such effects, a follow-up study is conducted by using a talking head as a speech synchronized visual support for an audio-based story.

4 Experiment 2: naturalness and eye-gaze behavior

To investigate the perceived non-verbal effect of synthesizing the gestures coupled with auditory prominence, an ecological experimental setup was designed by using the SynFace speech driven talking head [27] designed as a visual support for hard of hearing persons [37]. The talking head used in SynFace is the same talking head used in the previous experiment, but SynFace applies a real-time phoneme recognition on the input speech signal to guide the talking head, rather than forced alignment, as was the case in the previous experiment.

The experiment setup was based on presenting an audio novel (audio book) of 15 minutes with added noise, while the audio was supported visually using SynFace. The audio book was presented to moderately hearing impaired subjects (with a hearing sensitivity of 41-55 dB) and the subjects had to evaluate the talking head in different conditions. The audio book was divided into 3 parts, each 5 minutes long, and randomized into 3 conditions: one with audio-only audio book, one with the talking head with only articulatory movements, and the third one with the talking head visualizing gestures guided by automatically detected pitch accents. During listening to the audio book, the subjects' eye-gaze was tracked using a Tobii T120[1] eye-gaze tracker, integrated into the monitor on which the talking head was presented. Figure 4 shows a snapshot of a subject listening to the audio book and looking at a screen with a Tobii eye-gaze tracker. At the end of listening to each of the audio book conditions, the subjects were presented with a questionnaire form, Table 2 shows the questions included in the questionnaire form.

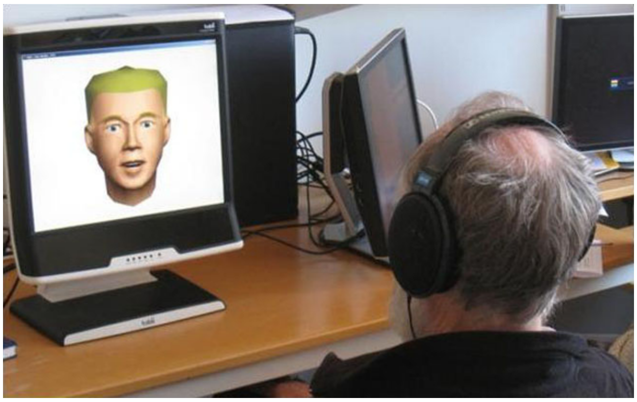

Fig. 4 A snapshot of a subject doing the audio book test with the talking head presented through the Tobii gaze tracker

Table 2 The format of the questionnaire form presented to the subjects

Question	Answers [1 .. 5]
1—How much did you understand?	nothing..everything
2—How effortful was it to understand?	no effort..much
3—Did the face help?	not at all..a lot
4—How much did you watch the face?	not at all..all time
5—How natural is the face?	unnatural..natural

This setup allowed us not only to measure the subjective opinion about the talking head, but also to study the behavioral change of the subjects eye gaze resulting from synthesizing non-verbal gestures to the talking head.

4.1 Stimuli preparation

Subjects firstly had the option to choose one of a long list of audio books obtained from the Swedish Library of Audio- and Braille books.[2] The first 15 minutes of each of the audio books chosen by the subjects were taken, and split into 3 parts, each of 5 minutes. The three parts were then randomly manipulated into three conditions: Only audio, audio and talking head (articulatory movements only), and audio and talking head (articulatory movements and gestures). Refer to Sect. 4.2 for a detailed explanation on how the gestures are generated.

After generating the stimuli, 6-speakers stereo babble noise was added to the audio signal to reduce the understanding level of the audio book. To decide the SNR level of the audio book, a small speech-in-noise intelligibility experiment was conducted for each subject to calibrate the SNR level.

[1] http://www.tobii.com/.

[2] The Library of Audio-and Braille books: http://www.tpb.se.

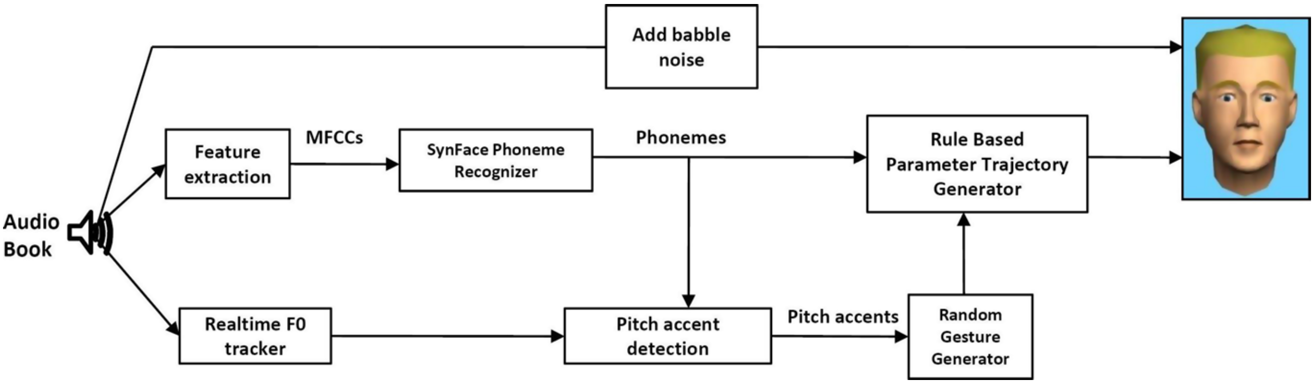

Fig. 5 Flow chart diagram of the talking head system used in the experiment, including noise and gestures addition

For the calibration experiments, 15 short Swedish everyday sentences were used. These sentences were between 7–9 syllables in length, and developed specifically for intelligibility testing, based on [2], and used in several previous intelligibility experiments (cf. [26]). Each of these sentences contains three words tagged as keywords (for scoring). The SNR was calibrated with a step of 1 dB starting initially at an SNR equal to 0 dB to reach an approximate of 50% keyword intelligibility rate. After calibration, the audio book SNR level was changed to the calibrated value for each subject plus 2 dB to allow for an easier understanding of the signal (which in principle should be supported further by viewing the talking head).

4.2 Gesture generation

For all the audio book parts, pitch accents were detected using the method described in Sect. 3.3.4. Since the automatic detection function used in Sect. 3.3.4 showed that it enhanced speech intelligibility significantly when coupled with an eyebrow raise, this gives the possibility of automatically detecting pitch accents which allows for a direct implementation in talking heads.

As to what gestures were to be synthesized, in [11], all facial parameters exhibited big variations under prominent words, and in the previous experiment, head nods, as well as eyebrow raise gestures, have been shown to aid speech comprehension. In addition to these cues, eye blinks have been used as a correlate to stress, and used in talking heads as in [38] based on a model suggested by P. Ekman in [39]. Eye blinks are important biological movements that have been suggested to synchronize with speech production [40]. They play an important role in the perception of natural facial movement, in speaking, reading and listening situations and are highly correlated with the cognitive state of the speaker [41].

To introduce variability to the facial movements in the talking head, head nods, eyebrow raise, and eye blinks are coupled with the automatically detected pitch accents.

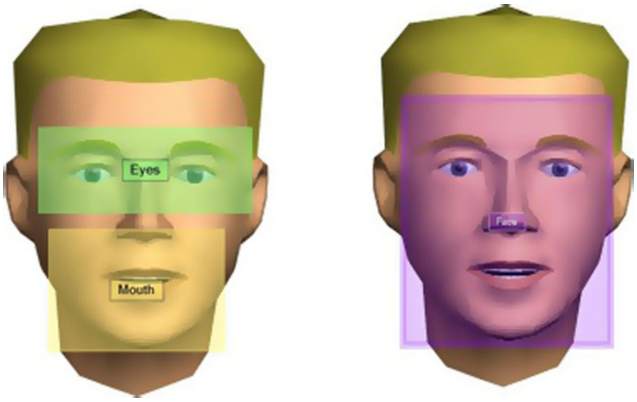

Fig. 6 *Left*: The eyes and mouth regions of the face. *Right*: The inside and outside regions of the face

It is important to note that gestures are highly variable, and correlations between certain acoustic events and facial movements are also highly variable [3]. In many applications, the occurrence of these gestures is modeled as a random process, since the factors which play a role in the decision of producing a certain acoustic gesture and facial gesture simultaneously depend on stochastic and less understood parameters (e.g. physical and cognitive status, attitudes, cultural background, etc.). Nonetheless, as the previous experiment showed, the meaning of the gestures might be perceived when they are synthesized almost arbitrarily.

In order to regulate the amount of gestures and avoid unnatural movement, a maximum of one gesture was allowed to be synthesized inside a moving window of 1 second, that is depending on the level of the pitch accents detected inside this window (see Sect. 3.3.4). After deciding on the place where the gesture will be added, a uniform random function chooses one of the head nods, eyebrow raise gestures, or blinks to be added.

Figure 5 displays the data flow for generating the audio book stimuli with gestures.

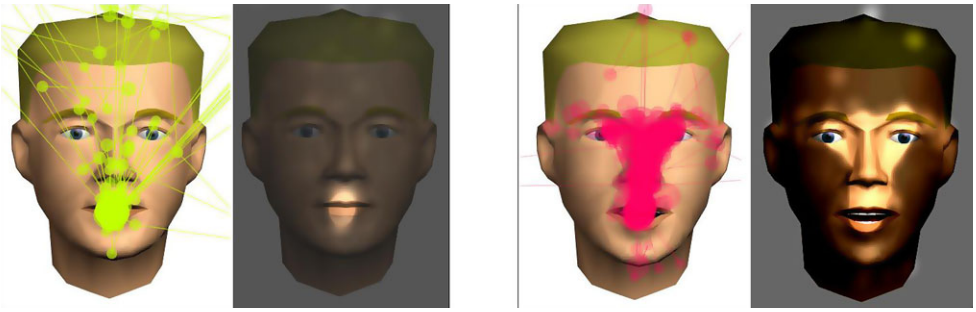

Fig. 7 Gaze tracking results for the ten subjects, showing gaze plot and heat map for both conditions (with and without gestures)

4.3 Subject and setup

A set of 10 moderately hearing impaired male subjects took part in the experiment, with an average age of 71.2 years old.

Subjects were equipped with Sennheizer HD-600 headphones and seated approximately 60 cm from a Tobii T120 screen with integrated eye-gaze tracker. The audio sampling rate of the acoustic stimuli was 16 kHz.

The experimenter was sitting next to the subjects during the whole experiment in order to administer the questionnaires. Figure 4 shows a snapshot of a subject listening to the audio book.

4.4 Analysis

4.4.1 Gaze analysis

The gaze tracking data of the audio book sessions was collected for all the 10 subjects, and divided into 2 parts for each subject: gaze for a talking head with gestures, and gaze for the talking head without gestures. The first test split the data into two Areas Of Interest (AOI): Face and Not Face, as displayed on the right in Fig. 6. The second test split the data into three AOI: Mouth region, Eyes region, and Other. Other represents gaze movements outside the mouth and eyes regions. The left side of Fig. 6 shows the location of these AOIs. Two types of visualizations are applied to present the gaze data, the first one is called a Gaze Plot which visualizes the fixations as circles and the saccades as lines between fixations. The size of the circles defines the length of the fixation, so the bigger the circle the longer the fixation. The other type is called Gaze Heat Map where the face is cleared in regions where there is higher intensity of gaze and shadowed on less visited areas, both visualizations are generated from the same data.

Figure 7 shows the Gaze Plot and Heat Map for all the subjects for the two SynFace versions (left: without gestures, right: with gestures). In this figure, it is clear that a wider area of the face was scanned by the subjects in the SynFace with gestures, specially around the eyes (eye blinks and eyebrow raising), this behavior is more similar to the normal eye-gaze behavior during audio-visual speech perception in noise that is reported in the literature (where no more than half the time is spent on the mouth, and alternations between the eyes and the mouth are frequent) [41, 42]. While in the talking head with no gestures condition (left), the subjects only focused on the mouth, which could be due to the understanding that information about the audio signal is only present in and around the lips and articulators. In the talking head with gestures condition, the subjects browsed the face at all time by fixations and saccades from the mouth to the eyes. It is also clear in the plot that the saccades into external regions outside the face have been significantly reduced with the use of gestures in the face.

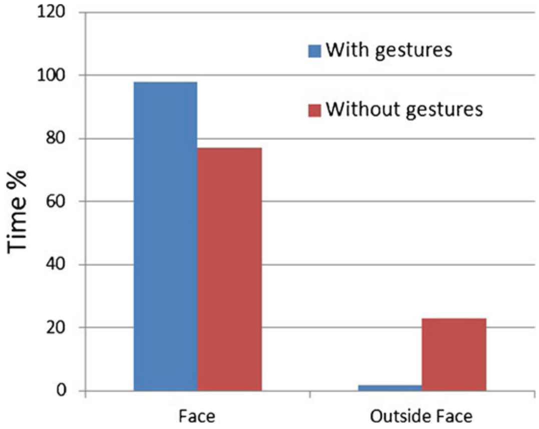

Fig. 8 The percentage of time spent inside and outside the face in the two conditions

To measure the time the user spent looking at the face during listening to the audio book, the sum of the fixations' length is measured for the areas of interest for all subjects, and for both versions of the face. Figure 8 presents the average percentage of time in and outside the face for both versions of the face. The figure shows that the time spent by the subjects looking at the face has increased from 76% to 96% when gestures are synthesized.

Figure 9 presents the average percentage of time on the defined regions of the eyes, mouth and the rest of the face. Comparing the two conditions, it can be noted that the eyes region has received significantly more gaze—and the area outside the face has received significantly less—in the "with gestures" condition than in the "without gestures" condition.

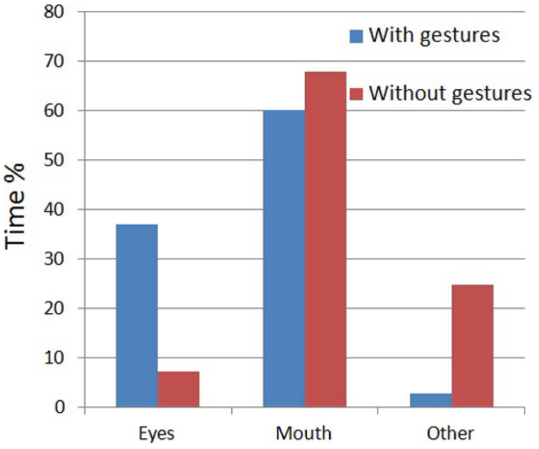

Fig. 9 The percentage of time spent on the eyes, mouth and the rest of the face in the two conditions

This can be explained as that subjects were more engaged with the talking head so that the time spent not looking at the face was lessened when gestures were synthesized. It is also interesting to see that the time spent looking at the mouth has lessened when gestures are synthesized, this means that when gestures are generally present on the face, they draw the gaze away from the mouth.

This, nonetheless, does not mean that the information perceived from the mouth region is reduced when gestures are synthesized since peripheral vision can still capture lip movements even if the eyes are not fixated on the lips. In a study in [43] on perceiving the McGurk effect, it is shown that audio visual speech perception is not influenced when the gaze was displaced from the talker's face, and that the perception of the effect is not affected under eccentric viewing conditions. Furthermore, they demonstrate that the analysis of high spatial frequency information afforded by direct oral foveation is not necessary for the successful processing of visual speech information.

4.4.2 *Questionnaires Analysis*

Looking at Table 3, the questions in the questionnaire target the subjective opinions of the test subjects after listening to the audio book for each version of the face.

Figure 10 shows the mean and standard deviation of the difference between the ratings for the *with gestures* face and the *no gestures* face for all the questions, hence the 0 point represents the rating of the *no gestures* condition for each of the questions.

An ANOVA was run with question rating and subject as dependent variables, and condition (with gestures, without gestures) as a dependent variable on each of the questions. The results show a significant mean effect for the Helpfulness (third question), Watching Duration (fourth question) and Naturalness (fifth question). Table 3 shows the degrees of freedom (df), the F value, and the *p* value for each of the tests.

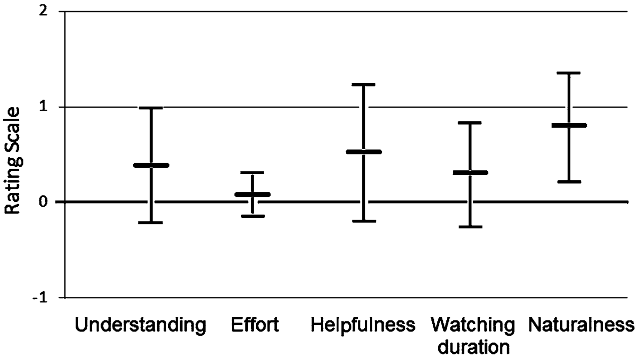

Fig. 10 A plot of the mean and standard deviation of the questionnaire ratings for the face with gestures conditions compared to 0 as the answer for the face with no gestures

Table 3 The ANOVA statistics on the answers of the questionnaires

Question	df	F	*p*
1 (Understandability)	1	4.31	0.06
2 (Effort)	1	0.59	0.46
3 (Helpfulness)	1	5	0.049
4 (Watching duration)	1	5.01	0.048
5 (Naturalness)	1	18.5	0.002

The results of the questionnaires show that, consistent with the results from the gaze data, that subjects spend more time watching the face when gestures are synthesized. Additionally, the type, rate, and timing of gestures over pitch accents result in an animated talking face which is significantly more natural, and subjectively presents a more understandable face than that with no gestures synchronized with pitch accents.

5 General discussion

The results from the intelligibility experiment indicate that when head nods and eyebrow raise gestures are visualized over prominent syllables, they can aid speech perception. On the other hand, the results do not indicate a strong evidence on whether visualizing them over non-prominent syllables may hinder or aid perception.

The speech signal in this test was degraded using a noise excited vocoder, and no pitch information was available to the listener, which may result in a decreased amount of information about the syllabic boundaries in the signal. The visualization of gestures then, might be a possible source of voicing information (which aligns with the significant increase of the condition with pitch accents (PA) over the condition with no gestures (N). We believe that an important function of gestures is temporal resolution and segmentation of the acoustic stream. If gestures over prominence provide information about the syllable boundaries of the prominent syllable, this can, in addition to providing semantic and pragmatic information about the sentence (the communicative function of prominence), it might provide segmental information (syllabification) of the underlying acoustic stream.

In Japanese [44], it was found that pitch accent can help in the selection of word candidates. In Swedish, syllables in words are contrasted through lexical stress. It is possible that visual prominence, aligned with prominent syllables, can provide information about the segmental structure of the underlying word, and hence help in shortening the candidate list for the mental lexicon access.

It was shown before that the perception of head movements can increase speech intelligibility [4], and that the motion at only the top of the head can do the same but more reliably in expressive sentences [5]. These studies have used, as stimuli, human recordings of head movements, and hence could not provide quantified information on when these movements communicated their effect. The present experiment, in addition to showing that visual cues of acoustic prominence can aid speech intelligibility, also describes this effect through the use of a minimal model of fixed head nods and eyebrow raise movements on well-defined instants in time.

In [45], there is neurophysiological evidence that matching visual information speeds up the neural processing of auditory information, although where and when the audio-visual representation of the audio-visual signal is created remains unsolved. It is evident that perceiving visual information increases the processing of the auditory stream and hence provides more temporal resolution; from this view, visualizing prominence may also provide information about speech rhythm and syllable boundaries of the underlying linguistic segment.

Going from the verbal effects of visual prominence to the interaction effects, we conducted the audio-book experiment. It has long been recognized that there is much information present in the face in addition to the articulators, which provide information about the speech. Taking these non-verbal cues into account makes animated characters more human-like by exhibiting more varying, complex and human-like behavior. In this study, we demonstrated that when gestures are coupled with prominence in an acoustically degraded stimuli, the gaze behavior of subjects significantly changed into patterns closer to those when looking at a real human face; moreover, this also lessened the time the subjects spent looking away from the face. This suggests that the subjects' engagement with the talking head is increased when these gestures are visualized, while these gestures added to the naturalness and helpfulness of the talking face. Many subjects have reported after the experiments that they did not notice that the talking head had embedded gestures in it, which might be a possible indication that the perception of these gestures is realized on a subconscious level; while other subjects have reported that they were excited by perceiving gestures manifested by the face, and that it was an indication of some sort of intelligence of the virtual agent.

6 Conclusion

We have investigated whether visual correlates to prominence can increase speech intelligibility. The experimental setup in this study used a lip synchronized talking head. By conducting an audio-visual speech intelligibility test, using facial gestures over prominent and pitch accented syllables, it was found that head nods and eyebrow raise gestures significantly increase the recognition rate. These results reveal new evidence that information about the verbal message is carried by non-verbal visual gestures. The experiment also provides a possibility to deploy these gestures in talking heads which would provide a medium for higher audio-visual speech perception. We also investigated the effects of synthesizing these gestures over pitch accented syllables on the eye-gaze behavior and on the subjective opinions of moderately hearing impaired subjects; the results show that users' eye gaze extends from only the mouth region in the articulatory only face to the eyes and mouth in the gestural face. In addition to that, the subjects opinions through questionnaires show an increase in intelligibility of the face when these gestures are added.

These results open the possibility for talking heads to use visual correlates to prominence to support visual speech perception and aid the communication of prominence through the facial modality. An important application is to synthesize these cues in talking avatars in speech-enabled multimodal user interfaces.

While the experiments show the roles of gestures synchronized with prominence, the question on how exactly users parse this audio-visual information and enhance their speech perception using it is still to be investigated. Another interesting question for further study concerns the optimal implementation of these gestures in terms of rate, design and amplitude for use in talking avatar systems.

Acknowledgements This work has been supported by the Swedish Research Council (grant #2005-3488), Stiftelsen Promobilia (grant #8038) and the Swedish national Graduate School for Language Technology—GSLT. We would like to thank all the subjects who took part in the experiments. Many thanks also to the anonymous reviewers for their valuable comments and feedback.

References

1. McGurk H, MacDonald J (1976) Hearing lips and seeing voices
2. Summerfield Q (1992) Lipreading and audio-visual speech perception. Philos Trans Biol Sci 335(1273):71–78

3. Cave C, Guaïtella I, Bertrand R, Santi S, Harlay F, Espesser R (1996) About the relationship between eyebrow movements and Fo variations. In: Proc of the fourth international conference on spoken language, vol 4
4. Munhall K, Jones J, Callan D, Kuratate T, Vatikiotis-Bateson E (2004) Head movement improves auditory speech perception. Psychol Sci 15(2):133–137
5. Davis C, Kim J (2006) Audio-visual speech perception off the top of the head. Cognition 100(3):21–31
6. Cvejic E, Kim J, Davis C (2010) Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. Speech Commun 52
7. Terken J (1991) Fundamental frequency and perceived prominence of accented syllables. J Acoust Soc Am 89:1768
8. Gundel J (1999) On different kinds of focus. In: Focus: linguistic, cognitive, and computational perspectives, pp 293–305
9. Grice M, Savino M (1997) Can pitch accent type convey information status in yes-no questions. In: Proc of the workshop sponsored by the association for computational linguistics, pp 29–38
10. Granström B, House D (2005) Audiovisual representation of prosody in expressive speech communication. Speech Commun 46(3–4):473–484
11. Beskow J, Granström B, House D (2006) Visual correlates to prominence in several expressive modes. In: Proc of the ninth international conference on spoken language processing
12. House D, Beskow J, Granström B (2001) Timing and interaction of visual cues for prominence in audiovisual speech perception. In: Proc of the seventh European conference on speech communication and technology
13. Swerts M, Krahmer E (2006) The importance of different facial areas for signalling visual prominence. In: Proc of the ninth international conference on spoken language processing
14. Krahmer E, Swerts M (2007) The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. J Mem Lang 57(3):396–414
15. Dohen M, Lœvenbruck H (2009) Interaction of audition and vision for the perception of prosodic contrastive focus. Lang Speech 52(2–3):177
16. Dohen M, Lcevenbruck H, Hill H (2009) Recognizing prosody from the lips: is it possible to extract prosodic focus. In: Visual speech recognition: lip segmentation and mapping, p 416
17. Streefkerk B, Pols L, Bosch L (1999) Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's. In: Sixth European conference on speech communication and technology, Citeseer
18. Fant G, Kruckenberg A, Nord L (1991) Durational correlates of stress in Swedish, French, and English. J Phon 19(3–4):351–365
19. Bruce G (1977) Swedish word accents in sentence perspective. LiberLäromedel/Gleerup, Malmo
20. Gussenhoven C, Bruce G (1999) Word prosody and intonation. In: Empirical approaches to language typology, pp 233–272
21. Heldner M, Strangert E (2001) Temporal effects of focus in Swedish. J Phon 29(3):329–361
22. Fant G, Kruckenberg A, Liljencrants J, Hertegård S (2000) Acoustic phonetic studies of prominence in Swedish. KTH TMH-QPSR 2(3):2000
23. Fant G, Kruckenberg A (1994) Notes on stress and word accent in Swedish. In: Proceedings of the international symposium on prosody, 18 September 1994, Yokohama, pp 2–3
24. Krahmer E, Swerts M (2004) More about brows: a cross-linguistic study via analysis-by-synthesis. In: From brows to trust: evaluating embodied conversational agents, pp 191–216
25. Massaro D (1998) Perceiving talking faces: from speech perception to a behavioral principle. MIT Press, Cambridge
26. Agelfors E, Beskow J, Dahlquist M, Granström B, Lundeberg M, Spens K-E, Öhman T (1998) Synthetic faces as a lipreading support. In: Proceedings of ICSLP'98
27. Salvi G, Beskow J, Al Moubayed S, Granström B (2009) Synface—speech-driven facial animation for virtual speech-reading support. J Audio Speech Music Process 2009
28. Beskow J (1995) Rule-based visual speech synthesis. In: Proc of the fourth European conference on speech communication and technology
29. Sjölander K (2003) An HMM-based system for automatic segmentation and alignment of speech. In: Proceedings of fonetik, pp 93–96
30. Beskow J (2004) Trainable articulatory control models for visual speech synthesis. Int J Speech Technol 7(4):335–349
31. Shannon R, Zeng F, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270(5234):303
32. Fant G, Kruckenberg A, Nord L (1991) Durational correlates of stress in Swedish, French and English. J Phon 19(1991):351–365
33. Heldner M, Strangert E (2001) Temporal effects of focus in Swedish. J Phon 29:329–361
34. Moubayed S Al, Ananthakrishnan G, Enflo L (2010) Automatic prominence classification in Swedish. In: Proceedings of prosodic prominence: perceptual and automatic identification workshop, Chicago, USA
35. Swerts M, Krahmer E (2004) Congruent and incongruent audiovisual cues to prominence. In: Proc of speech prosody
36. de Cheveigne A, Kawahara H (2002) YIN, a fundamental frequency estimator for speech and musicy. J Acoust Soc Am 111:1917
37. Al Moubayed S, Beskow J, Oster A-M, Salvi G, Granström B, van Son N, Ormel E (2009) Virtual speech reading support for hard of hearing in a domestic multi-media setting. In: Proceedings of interspeech 2009
38. Poggi I, Pelachaud C, De Rosisc F (2000) Eye communication in a conversational 3D synthetic agent. AI Commun 13(3):169–181
39. Ekman P (1979) About brows: Emotional and conversational signals. In: Human ethology: claims and limits of a new discipline: contributions to the colloquium, pp 169–248
40. Cassell J, Pelachaud C, Badler N, Steedman M, Achorn B, Becket T, Douville B, Prevost S, Stone M (1994) Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In: Proceedings of the 21st annual conference on computer graphics and interactive techniques, pp 413–420
41. Raidt S, Bailly G, Elisei F (2007) Analyzing and modeling gaze during face-to-face interaction. In: Proceedings of the international conference on auditory-visual speech processing (AVSP 2007)
42. Vatikiotis-Bateson E, Eigsti I, Yano S, Munhall K (1998) Eye movement of perceivers during audiovisual speech perception. Percept Psychophys 60(6):926–940
43. Paré M, Richler R, ten Hove M, Munhall K (2003) Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect. Percept Psychophys 65(4):553
44. Cutler A, Otake T (1999) Pitch accent in spoken-word recognition in Japanese. J Acoust Soc Am 105:1877
45. van Wassenhove V, Grant K, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Nat Acad Sci 102(4):1181