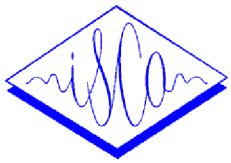
ISCA Archive
http://www.isca-speech.org/archive



AVSP 2009 - International Conference
on Audio-Visual Speech Processing
University of East Anglia, Norwich, UK
September 10-13, 2009

Effects of Visual Prominence Cues on Speech Intelligibility

Samer Al Moubayed and Jonas Beskow

KTH Centre for Speech Technology, Stockholm, Sweden.
{sameram,beskow}@kth.se

Abstract

This study reports experimental results on the effect of visual prominence, presented as gestures, on speech intelligibility. 30 acoustically vocoded sentences, permutated into different gestural conditions were presented audio-visually to 12 subjects. The analysis of correct word recognition shows a significant increase in intelligibility when focally-accented (prominent) words are supplemented with head-nods or with eye-brow raise gestures. The paper also examines coupling other acoustic phenomena to brow-raise gestures. As a result, the paper introduces new evidence on the ability of the non-verbal movements in the visual modality to support audio-visual speech perception.
Index Terms: prominence, head-nod, eye-brow, speech intelligibility, talking heads, lip-reading, gesture, visual prosody.

1 Introduction

Recently, there has been an increasing interest in the verbal and non-verbal interaction between the visual and the acoustic modalities from a production and perception perspectives . Studies have reported possible correlations between acoustical prosody and certain facial movements. In [1] , correlation between f0 and eye-brows movements was discussed. In [2], correlations between f0 movements and head-movements 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 usually is highly variable, but an understanding of the redundancy of information in these two modalities can greatly help in, by exploiting it, developing audio-visual human-human and human-machine interfaces to guarantee maximum amount of interaction [3].

One of the prosodic phenomena which attracted much focus is prominence. Prominence is typically defined as when one linguistic segment is made salient in its context. Words (or longer or shorter linguistic segments) can be made prominent to convey information such as focus [4], and information status [5]. Hence, the communication of prominence can impact the interpretation of a word or phrase, and affect the speech comprehension [6].

Recent studies have focused on the relation between the visual modality (the face) and the acoustic prominence[3]. In [7], results on Swedish showed that in all expressive modes, words which receive a focal accent exhibit greater variation in the facial parameters movements (articulators, eye-brows, head, etc.) than when the word is in a non-focused position. In [8], visualizing eye-brows movements and head nods on a talking head is found to be a powerful cue to enforce the perception of prominence. In a study in [9], an investigation on the interaction between the acoustic and the visual cues of prominence was carried out, the result of this study revealed that, from a production perspective, when a word is produced with a visual gesture, the word received higher acoustic emphasis. It also suggests that, from a perception perspective, when people see a visual gesture over a word, the acoustic perception of the word' prominence is increased.

Since these studies support the strong relation between the auditory and the visual modalities in perceiving prominence. The important question is: Can visualizing prominence (as facial gestures) increase the speech intelligibility when its acoustic counterpart is absent or distorted?

This paper investigates this question, by conducting a speech intelligibility experiment in Swedish with the help of a lip-synchronized talking head.

The paper is organized as follows: Section 2 presents the method used, the design of the stimuli, and the participants. Section 3 presents the results of the experiments. Section 4 discusses the different results and the possible implications of the findings and Section 5 concludes the paper and suggests future works.

2 Method

2.1 Setup and Data

Computer synthesized talking heads have been progressively developing, offering the possibilities for many experimental designs which were not possible before. That is by manipulating and changing the stimuli in one modality while keeping the other modality intact, allowing for stimuli setup which explores the effects of specific variables, for example, by manipulating the required variable to measure, and keeping the others static [10]. This was the main reason to use a lip-synchronized talking agent as the medium for the visual modality in this work.

This experiment design deploys the approach of presenting human subjects with a vocoded speech signal, while looking at a rule based parametric talking head [11]. Different sets of sentences will receive facial gestures at different timings along the speech signal, and the difference in cross-subject speech intelligibility is to be 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 speech synthesis voice creation.

The speech files of the 40 sentences were force-aligned using an HMM aligner [12] to guide the talking head lips movement. The audio was processed using a 4-channel noise excited vocoder [13] to reduce intelligibility. The number of channels was decided after a pilot test to ensure that an intelligibility rate between 25% and 75%, that is to avoid any upper and lower 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 vocoding used. The 30 sentences left 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 (detailed in section 2.3). These groups were systematically permutated among 12 normal hearing subjects (with 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 could understand from the signal.

2.2 Prominence Marking

In the Swedish intonation model [14], focal-accent is the highest level of prominence. Focal accent is a word level perceptual category. According to the Swedish prominence model, the acoustic correlates to focal-accent can be distributed over more than one syllable in a word, that is by having a word accent fall, followed later by an accent rise. This is more evident in poly-syllabic words (eg. 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 [15][16].

For the purpose of this study, since 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 [8]), we decided to limit the size of the focused segment from the whole focused word to its most prominent syllable.

To establish that, one native Swedish speech expert had listened to all the 30 test sentences, and marked them temporally with prominence. By investigating the prominence markers, all sentences have received between 1 to 3 prominence marks, and the overall number of marks in the 30 sentences summed to 60.

2.3 Stimuli Conditions

As mentioned before, every sentence in the test set was played back in 6 different visual variants. Following is a detailed description of 5 of these variants (the sixth condition was a special purpose variant and is left out of this analysis).

It is also important to mention here, that whenever a sentence has 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 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.

2.3.1 No Gesture (N)

The first condition was 'articulators-only' where the face looked 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 is aimed to be a control measurement for the rest of the conditions. Figure 1a displays the talking head in the neutral position.

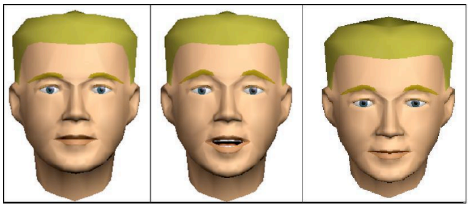


Figure 1: Snapshots of the talking head in different gestural positions. (a) neutral parameters. (b) peak of the eye-brows raise gesture. (c) peak of the head-nod .

2.3.2 Head-nods (H)

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 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 estimation of the average length of a stressed syllable in Swedish. Figure 1c shows the talking head at the lower turning point of the head nod.

2.3.3 Eye-brows Raise (EB)

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 1b shows the eye-brow gesture at its top turning point.

2.3.4 Pitch Slopes Eye-brows raise (P)

A perception experiment in Dutch [17], 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. In addition, as we mentioned earlier, other studies have suggested a possible correlation between f0 and eye-brows movements.

In this condition, eye-brow 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. 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).

2.3.5 Random Eye-brows 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, 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. words or syllables 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 gestures overlap.

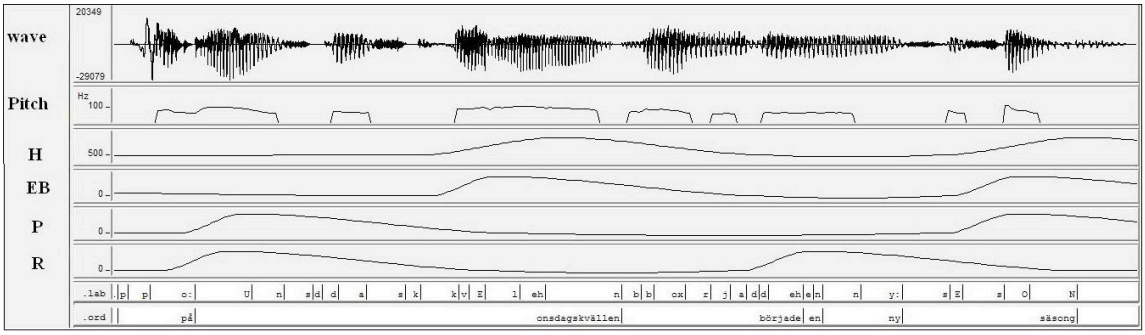


Figure 2: An example sentence ”På onsdagskvällen började en ny säsong”, ”On the Wednesday evening a new season began”, shown with the different head or eye-brows movements according to the condition.

3 Results

A percent-correct word recognition scoring was applied to the answer of each of the sentences of the subjects. As a result, every condition contained 60 samples (12 subject * 5 sentences per condition). Since subjects differed in their average recognition rate (different abilities to understand the vocoded signal), the answers of each subject were normalized to a standard distribution. A two-sample t-test was applied to the samples of every two conditions. Table 1 presents the p-value and 95% confidence interval for each couple of conditions. Figure 3 shows the boxplot of the samples per condition. Looking at the significance values, there is a clear significant increase in word recognition rate for the head-nod (H), eye-brows raise (EB) and pitch-slopes eye-brows (P) conditions over the no-gesture (N) condition. There was no significant difference in between the H, EB, and P conditions. The random eye-brows condition (R), although had an increase in the mean value, had no significant mean difference from the no-gesture (N) condition. The only measure which had significant difference from the random condition (R) was the head-nod (H) condition, which, in terms of mean recognition rate, had the highest value among all the other conditions (Figure 3).

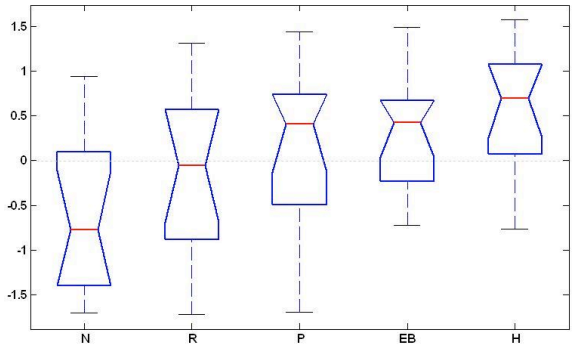


Figure 3: Box plot of the % correct word recognition over conditions (normalized over subjects).

4 Discussion

The results of this experiment indicate that when head-nod and eye-brow 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 weather visualizing them over non-prominent syllables may hinder or aid the perception. The speech signal in this test was vocoded, 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 P condition over the no-gesture N condition).

Previously, head nods have been shown to be a stronger cue in the perception of prominence over eye-brows [8], which seems to be in line with the results in this experiment (mean increase, but not significant difference). Head nods might perhaps be a stronger indication of prominence perceptually because of their larger area in surface motion, and hence requiring less cognitive effort to realize compared to the perception of eye-brows movements which are realized locally, separately from the the lips movements. However, it is hard to investigate from this experiment, in what way the visual realization of prominence has aided the speech perception task.

In Japanese [18], 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

Table 1: Results from the multiple comparison test. Tables shows every two samples (conditions) and their p-value and a 95% confidence interval.

Condition	p-value	95% CI
N*H	<0.001	[-1.8869 -0.5937]
N*EB	<0.005	[-1.5879 -0.3148]
N*P	<0.05	[-1.5708 -0.0896]
H*R	<0.02	[-1.4795 -0.0996]
N*R	>0.2	[-1.2028 0.3012]
H*EB	>0.2	[-0.8508 0.2730]
H*P	>0.2	[-1.0876 0.2674]
EB*P	>0.2	[-0.5467 0.7891]
EB*R	>0.1	[-0.1799 1.1812]
P*R	>0.2	[-1.2502 0.4280]

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 [2], and that the motion at only the top of the head can do the same but more reliably in expressive sentences [19]. 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 quantifies this effect through the use of a minimal model of fixed head nods and eye-brows raise movements on well-defined instats in time.

It's important to stress that this study does *not* claim that these movements (head nods and eye-brows gestures) are in any way optimal or identical to movements employed by humans to communicate prominence through head and eye-brows (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 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 eye-brows movements, but it shows that these cues can be of help to speech intelligibility when the acoustic signal is degraded.

5 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 syllables, it was found that head nods and eye-brows raise gestures significantly increase the recognition rate. This result opens 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 of these findings is for the implementation of talking-head-based visual speech support systems for the hard-of-hearing, such as SynFace [20]. To facilitate this, one current research direction is the development of automatic acoustic prominence detection systems for the purpose of driving facial gestures.

6 Acknowledgement

This work is supported by the European Commission project H@H-Hearing at Home (IST-045089). We would like to thank all the participants for their willing to participate in the experiment.

References

[1] C. Cave, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, "About the relationship between eyebrow movements and Fo variations," in *Proc of the Fourth International Conference on Spoken Language*, vol. 4, 1996.

[2] K. Munhall, J. Jones, D. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Head Movement Improves Auditory Speech Perception," *Psychological Science*, vol. 15, no. 2, pp. 133–137, 2004.

[3] B. Granström and D. House, "Audiovisual representation of prosody in expressive speech communication," *Speech Communication*, vol. 46, no. 3-4, pp. 473–484, 2005.

[4] J. Gundel, "On different kinds of focus," *Focus: Linguistic, cognitive, and computational perspectives*, pp. 293–305, 1999.

[5] M. Grice and M. Savino, "Can pitch accent type convey information status in yes-no questions," in *Proc of the Workshop Sponsored by the Association for Computational Linguistics*, 1997, pp. 29–38.

[6] P. Keating, M. Baroni, S. Mattys, R. Scarborough, A. Alwan, E. Auer, and L. Bernstein, "Optical phonetics and visual perception of lexical and phrasal stress in English," *Proc of the Fiftheenth International Conference on Spoken Language*, pp. 2071–2074, 2003.

[7] J. Beskow, B. Granström, and D. House, "Visual correlates to prominence in several expressive modes," in *Proc of the Ninth International Conference on Spoken Language Processing*, 2006.

[8] D. House, J. Beskow, and B. Granström, "Timing and interaction of visual cues for prominence in audiovisual speech perception," in *Proc of the Seventh European Conference on Speech Communication and Technology*, 2001.

[9] M. Swerts and E. Krahmer, "The importance of different facial areas for signalling visual prominence," in *Proc of the Ninth International Conference on Spoken Language Processing*, 2006.

[10] D. Massaro, *Perceiving talking faces: From speech perception to a behavioral principle*. Mit Press, 1998.

[11] J. Beskow, "Rule-based visual speech synthesis," in *Proc of the Fourth European Conference on Speech Communication and Technology*, 1995.

[12] K. Sjölander, "An HMM-based system for automatic segmentation and alignment of speech," in *Proceedings of Fonetik*, 2003, pp. 93–96.

[13] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, p. 303, 1995.

[14] G. Bruce, *Swedish word accents in sentence perspective*. LiberLäromedel/Gleerup, 1977.

[15] G. Fant, A. Kruckenberg, and L. Nord, "Durational correlates of stress in Swedish, French and English," *Journal of Phonetics*, vol. 19, no. 1991, pp. 351–365, 1991.

[16] M. Heldner and E. Strangert, "Temporal e4ects of focus in Swedish," *Journal of Phonetics*, vol. 29, pp. 329–361, 2001.

[17] M. Swerts and E. Krahmer, "Congruent and incongruent audiovisual cues to prominence," in *Proc of Speech Prosody*, 2004.

[18] A. Cutler and T. Otake, "Pitch accent in spoken-word recognition in Japanese," *The Journal of the Acoustical Society of America*, vol. 105, p. 1877, 1999.

[19] C. Davis and J. Kim, "Audio-visual speech perception off the top of the head," *Cognition*, vol. 100, no. 3, pp. 21–31, 2006.

[20] J. Beskow, I. Karlsson, J. Kewley, and G. Salvi, *SYNFACE - A talking head telephone for the hearing-impaired*. Springer-Verlag, 2004, pp. 1178–1186. [Online]. Available: http://www.speech.kth.se/prod/publications/files/1065.pdf