Multimodal perception of prominent words: the role of head nods and intonation in speech

Group members: Beata von Grothusen, Emma Igelström, Linnéa Lennartsson, Melissa Obrou

Supervisors: David House and Patrik Jonell

DM 2350 Human Perception for Information Technology

2018-10-17

Abstract

This study investigates whether head nods affect human perception of intonation of a word in a sentence. When speaking there are acoustic cues (word intonation) and visual cues (head nods) to relay emphasis of important words (Krahmer and Swerts, 2007). These work together to form the communication. Can the perception of the emphasised word in a sentence be changed/moved through making a head nod? The method used testing this hypothesis was to let 23 test subjects watch a film with associated audio. The film contained 36 short clips of an animation of a head that enunciates three different sentences, each with the head nod and the emphasised word on different places in each clip. The results of the study show that when the head nod was on the emphasised word almost all test subjects perceived that word as intoned. Whereas, when the head nod was moved to a different place a part of the test subjects perceive the word with the head nod as emphasised. This pattern was stronger in the weakly intoned words. Although, a majority still perceived the intoned word as emphasised. Thus indeed it is possible to move the perception of the emphasised word with a head nod when it is weakly intoned, but only in some cases.

Introduction

It is commonly known that word intonation is important for conveying information, also how body language is important in communication. Speakers may use both acoustic cues (word intonation) and visual cues (head nods) to emphasise the importance of a word (Krahmer and Swerts, 2007). The interesting thing is how do these two, sound and body, work together in communication and how do they affect the perceived communication?

Asor writes that research has found that head nods provide strong cues for the perception of speech especially when synchronized with the stressed vowel in Swedish. In particular for word recognition and prominence. Another study found that head nods are stronger cues of perception than eyebrow raising (Asor, 2014). Al Moubayed and Beskow also conclude that when it comes to perception seeing a visual gesture you perceive the vocal prominence of the word as strengthened. The study is however, inconclusive on the fact if a malplaced headnod or gestures hinders the perception of nearby prominent parts. The study is additionally inconclusive of why the visual realisation affects the speech perception (Al Moubayed and Beskow, 2009). All our background research, Asor, Kramer and Swerts, and Al Moubayed and Beskow, allude that visual cues have an effect on co-occurring speech. But whether malplaced visual cues has an effect as well is not as much researched. Thus, the studies in this article will test if a malplaced head nod can move the perception of the intonation.

Aim

The purpose of this research is to determine whether head nods affect human's perception of intonation of a word in a sentence. The question for this research that we will aim to answer is "Can the perception of the emphasised word in a sentence be changed/moved through making a head nod?". The aim will be to determine if performing a head nod on a word that is not emphasized in a sentence will be perceived as the emphasized word. The hypothesis is that it will be possible to change perception if the intonation is weak for the emphasised word and not possible with a strong intonation of the emphasised word. No attempts will be done to test how other types of body language, except head nods, affect perception of word intonation.

Method

To test the hypothesis that a head nod can change the perception of the emphasized word, 23 test subjects were used. Each test subject got to see a film (which was the same for everybody) of 9 minutes. The film was created from an animation interface (see Figure 3) created by Patrik Jonell in the 2D3D animation tool Unity. To make the animated face to be as realistic as possible, Emma, one of the group members, acted model to get all the mouth and eye movements (see Figure 1). The animation was recorded with an iPhone X, and Unity was then used again to create the head nods. This was done through the section for "JointShouldersMiddle - rotation", and the x, y, and z axis were changed so that all the movements from filming Emma disappeared.

Figure 1: The recording setup. Emma in front of an iPhone X. Computer open in Unity.

The only movements that were left were from the eyes and mouth. Points were put out in the animation exactly where the head nod was wanted. The range and speed of the head nod could then be changed with the points in the animation. The length of the head nods were adapted to the length of the word that contained the nod. Therefor making some head nods faster than others, depending on how short the word was. The amplitude of all the nods were the same. This was a conscious choice since we were not studying how different sized head nods affects perception. Different sized head nods were tested until we decided on one were it looked like a natural head nod. Not too exaggerated, but also not too discrete.

The subjects in the study were all KTH students. They were in the age of 20-35 and had no hearing or sight impairments. The tests were carried out in a room where the test

subject sat alone by a table, facing the wall, with a laptop and headphones (see Figure 2).

Figure 2: Test subject in the test setting

Some tests were carried out with two test subjects at the same time in the room with separate laptops. A brief instruction was given beforehand to the test subjects, that they would view a certain number of videos and listen for the word they perceived as emphasised. The test subjects were also instructed to both listen and watch the video.

The test subjects got to watch and listen to the film where three sentences were presented 12 times each in randomized order(see Appendix A). The three different sentences that were used are; "*Det är fint väder idag*", "*Hunden åt min läxa*", "*Solen skiner idag*". Each sentence was emphasized on two words; six times each, half the time the word was pronounced with a strong intonation and the other half with a soft intonation (see figure 3). The same audio was used for all the sentences were the intonation was strong and another for when the intonation was soft. The head nod was placed on a different word every time the test person was seeing and listening to each sentence. This was to test if the nod affected the test person to select the word that was being intonated or the word that had the head nod.

Audio 1	Audio 3
Film 1: **Hunden** åt min läxa	Film 7: Hunden åt min **läxa**
Film 2: **Hunden** åt min läxa	Film 8: Hunden åt min **läxa**
Film 3: **Hunden** åt min läxa	Film 9: Hunden åt min **läxa**
Audio 2	Audio 4
Film 4: ***Hunden*** åt min läxa	Film 10: Hunden åt min ***läxa***
Film 5: ***Hunden*** åt min läxa	Film 11: Hunden åt min ***läxa***
Film 6: ***Hunden*** åt min läxa	Film 12: Hunden åt min ***läxa***

Figure 3: One of the sentences that were used were "Hunden åt min läxa". The emphasised word is ***bold****, the nod is on the underlined word. When the word is* ***italic****[it] and* ***bold****[it], the word is having a weak emphasising.*

To make the test as thorough and with as little risk for interference as possible, the test subject got a paper with all the 36 sentences printed out to mark the word they felt was emphasised. This instead of having to change screens on the laptop. To help the test subject understand the task, they got to see a test version with a sentence that was not used later in the test. The test subject then got to see the 36 actual sentences with 10 seconds in between to mark which word they thought was intoned before a new sentence was presented on the screen. The 36 sentences were all in one long video so there was no interference from changing film or mixing with the computer.

Figure 4: Screenshots from video 12: "Hunden åt min läxa". With head nod on läxa.

Results

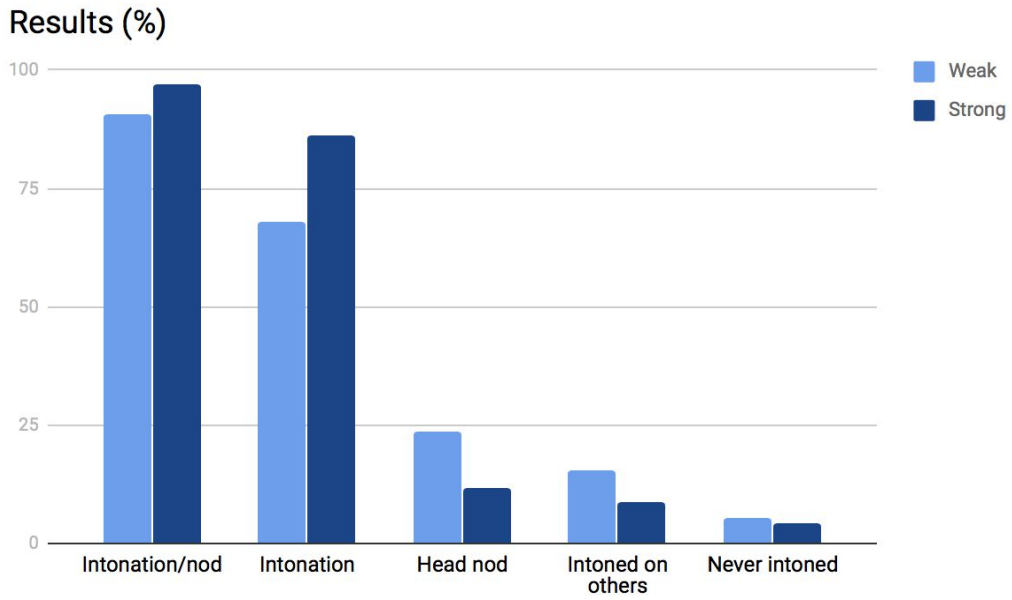

Figure 5: The result from the test, showing the percentage of respondents answering with different combinations of intonations and head nods. Intonation means that the head nod was on a different word, head nod means the intonation was on a different word, intoned on others means the answer was not intoned but had been intoned in a different film previously played and never intoned means the word is never intoned in that sentence during any of the 36 films.

The results were analysed through counting how many responds were written on every word in each of the 36 videos. An average score for each category (Intonation/nod, Intonation, Head nod, Intoned on others, never intoned) was then received through adding the answers in each category together and dividing by the tot
al number of videos (see Figure 4).

97.1% answered correctly (the intoned word) when the head nod and the strong intonation was on the same word. In the same situation, but with the intonation being weak, the result was 90.6%. When a word was strongly emphasised, but the head nod was on a different word, 86.2% still chose the intoned word. When the intonation was weaker and the head nod was on a different word, the result came to only 68%. When the head nod was on a different word than the strongly intoned word 11.7% answered that they perceived the nodded word as emphasised, but when the intonation was weaker this number increased to 23.7%. When a word had been emphasised in any of the other videos shown to the test subject and the same sentence was played again, the same word that had been intoned before was chosen in 8.7% of the cases with strongly emphasised sentences and 15.5% in weak cases. 4.3% in strong intonation cases, and 5.2% in weak intonation cases, people chose words that were never intoned or head nodded in any of the videos.

Discussion

To begin with, most subjects chose the rightly intoned word when the intonation and head nod was on the same word, both with weak and strong intonations. In fact, the right word was chosen in most cases of strong intonation. The results also show that more people chose the right word when the intonation was strong and were more prone to choose a word with the head nod when the intonation was weaker. This implies that the most prominent word in a sentence actually can be switched with a head nod, at least if the intonation on the word is weak. In addition to this, the results show that some test subjects were more prone to choose words that had been emphasised in the sentence previously during the test, especially if the intonation was weak. This might have affected the results in the way that the subject had an expectation of what word would be emphasised in the sentence and therefore chose that one without second thought.

One important thing to have in mind is that people can have different ways of perceiving word intonation. Some people consider a higher pitch to be emphasis while others consider lower pitch to be emphasis. When people listen to a conversation and try to interpret it, they listen for if the tone is higher or lower or if the tone is rising or falling. They do this by focusing on the final portion of the syllable (Xu and Wang, 2001). In the tests a higher pitch was used for intonation. One of the test persons commented on that she was used to interpret a lower pitch word to be the emphasised word. Therefore she first got confused by what words that had the intonation in the test. This can have affected the results for this research and it is of value

to be aware of this. Additionally, to have in mind is that the response time to an unfamiliar accent can be slower. A study made by Adank et. al. (2009) showed that standard English speakers had a slight delay in response time when listening to Glaswegian English speakers. While Glaswegian English speakers did not have a delay in response time when listening to standard English. Most likely due to the fact that they have been exposed to more standard English than the other group have been exposed to Glaswegian English. Thus, if the test subjects were unfamiliar with Emma's accent it could have affected their responses.

Since a real life person's voice was used for the speaking part of the test it is also important to be aware of that it may not be a perfect sentence. With that meaning the sentence having a straight baseline. Some words may be a little emphasised even though they should not have been. If doing this test again it could be a good idea to use a speech synthesis. In this way all the words could have the exact same tone, except the emphasised word. However, a computer voice is not as realistic as a human voice, and thus could have negative effects on the results as well since people are more used to interpreting real human voices. Another way to improve this could be by having a pilot study first where the test subjects in the pilot study evaluate the sentences before they are used in the study.

A strength of the study was the head nod animation. This made it possible to get a perfect head nod without any distractions or other movements of the head. Using a real life person might have given further distractions and not provide this perfect head nod that we aimed for to make it very clear when a head nod appeared, even though a real life person could have given a more clear mouthing of the word. We have not considered this as a factor that could have affected the results since the focus of this study was the head nod and the intonation.

The teamwork in this group has worked well and all members have had the same vision of what we want to achieve. This has contributed to a good work environment and a lot of work has been done during each meeting. This has lead to less stress and therefore a better creative process. One thing that could have lead to some disturbance is that every member of the group has not been able to attend on every meeting. But when absent form a meeting all members have worked from home and kept oneself up to date. So that never became a problem.

A clear timeplan was made in the beginning of this project and it have been followed. This has lead to that we have not fallen behind with our work. We have also decided when to meet for every week and therefore less confusion arose and a good mood in the group could be uphold. This is one major learning outcome for this group, that planning helps the workflow and also to keep a good group atmosphere.

Conclusion

This study shows that the hypothesis that head nods can change what word is perceived

as the most important word in a sentence, if the intonation of the word is weak, is actually true. The word most test subjects perceived as the important word was equal to the emphasised word when the intonation was strong, but more test subjects chose the word with head nod when the intonation was weaker. Since this study only was conducted in the Swedish language, it could be interesting to do similar studies and see if the results are different in other languages and also how cultural differences play a role in the perception of language. Similar studies on other types of body language and how they affect intonation in speech could also be interesting

References

Adank, Patti, Bronwen G. Evans, Jane Stuart-Smith, Sophie K. Scott. *Comprehension of Familiar and Unfamiliar Native Accents Under Adverse Listening Conditions. Journal of Experimental Psychology: Human Perception and Performance 2009, Vol. 35, No. 2, 520-529.* https://pdfs.semanticscholar.org/6e1c/581bcbe2b9665795e00a7754bed514fd79aa.pdf

Al Moubayed, S., J. Beskow (2009). *Effects of Visual Prominence Cues on Speech Intelligibility.* KTH Centre for Speech Technology, Stockholm, Sweden.

http://www.speech.kth.se/prod/publications/files/3357.pdf

Asor, E. *The timing of head nods is constrained by prosodic structure.* Universitat Pompeu Fabra, Barcelona, Spain.

https://repositori.upf.edu/bitstream/handle/10230/23429/Asor_2014.pdf

Krahmer, E., M. Swerts (2007). *The effects of visual beats on prosodic prominence: Acoustic analysis, auditory perception and visual perception.* Journal of Memory and Language, volume 57, issue 3, October 2007, p. 396-414.

https://ac.els-cdn.com/S0749596X07000708/1-s2.0-S0749596X07000708-main.pdf?_tid=d2b52974-a0e5-4321-a4d5-c8aae09d5293&acdnat=1537172525_5b5a9e1ca2069e67f604a2be8d2ab51b

Xu, Y., Q. E. Wang. *What Can Tone studies Tell us about Intonation?* Intonation: Theory, Models and Applications, Proceedings of an ESCA Workshop. European Speech Communication Association (A. Botinis, G. Kouroupetroglou, & G. Carayannis, editors), pp. 337-340. Athens, Greece: European.

http://www.haskins.yale.edu/yixu/Xu_ESCA97.pdf

Appendix A

Key:

Underline means headnod.

Bold means strong intonation.

Bold and italics means weak intonation.

Film 1: **Hunden** åt min läxa

Film 2: Det är **fint** väder idag

Film 3: Solen **skiner** idag

Film 4: Solen ***skiner*** idag

Film 5: Det är ***fint*** väder idag

Film 6: ***Hunden*** åt min läxa

Film 7: ***Hunden*** åt min läxa

Film 8: Solen skiner **idag**

Film 9: Hunden åt min **läxa**

Film 10: Hunden åt min ***läxa***

Film 11: Solen skiner **idag**

Film 12: **Hunden** åt min läxa

Film 13: Solen skiner **idag**

Film 14: Det är ***fint*** väder idag

Film 15: Hunden åt min **läxa**

Film 16: Hunden åt min ***läxa***

Film 17: Hunden åt min **läxa**

Film 18: ***Hunden*** åt min läxa

Film 19: Solen ***skiner*** idag

Film 20: Solen skiner ***idag***

Film 21: Det är ***fint*** väder idag

Film 22: Det är **fint** väder idag

Film 23: Solen **skiner** idag

Film 24: Solen ***skiner*** idag

Film 25: Det är fint **väder** idag

Film 26: Hunden åt min ***läxa***

Film 27: Det är fint ***väder*** idag

Film 28: Det är fint **väder** idag

Film 29: Solen skiner ***idag***

Film 30: Solen **skiner** idag

Film 31: Det är fint ***väder*** idag

Film 32: Det är fint ***väder*** idag

Film 33: **Hunden** åt min läxa

Film 34: Det är fint **väder** idag

Film 35: Det är **fint** väder idag

Film 36: Solen skiner **idag**