Can visualization of internal articulators support speech perception?

Preben Wik, Olov Engwall

Centre for Speech technology, School of Computer Science and Communication, KTH, Sweden
preben@speech.kth.se, engwall@kth.se

Abstract

This paper describes the contribution to speech perception given by animations of intra-oral articulations. 18 subjects were asked to identify the words in acoustically degraded sentences in three different presentation modes: acoustic signal only, audiovisual with a front view of a synthetic face and an audiovisual with both front face view and a side view, where tongue movements were visible by making parts of the cheek transparent. The augmented reality side-view did not help subjects perform better overall than with the front view only, but it seems to have been beneficial for the perception of palatal plosives, liquids and rhotics, especially in clusters. The results indicate that it cannot be expected that intra-oral animations support speech perception in general, but that information on some articulatory features can be extracted. Animations of tongue movements have hence more potential for use in computer-assisted pronunciation and perception training than as a communication aid for the hearing-impaired.

Index Terms: talking head, speech perception, speech visualization, audiovisual speech, internal articulation

1. Introduction

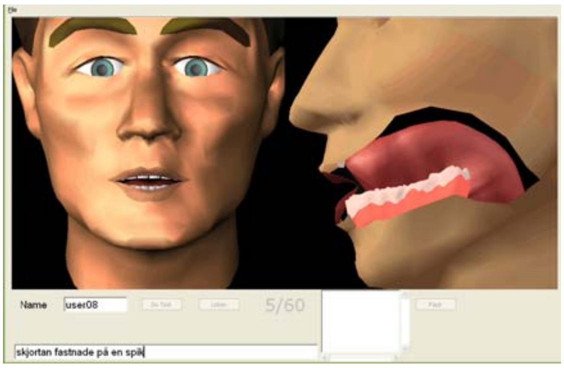
It is well established that visual information support speech perception, especially if the acoustic signal is degraded [1]. Hearing-impaired listeners can make use of speech reading abilities to better understand what is being said when they are face-to-face with the speaker. Normal-hearing listeners also benefit from information given by the face when the level of the background noise is close to that of the speech signal. It has been shown, e.g. in [2], that this gain is not only provided by a natural face, but also by synthetic faces.

Many phonemes are however impossible to identify by looking at the speaker's face, since the articulation of the tongue cannot be seen when the place of articulation is too far back. A growing community of hearing-impaired persons with residual hearing rely on cued speech [3] to supplement the acoustic signal and speech reading of the face. With cued speech, additional phonetic information is conveyed with hand sign gestures. These gestures are however an arbitrary iconic system that needs to be learnt and it may be more effective to display all relevant articulatory features faithful to actual speech production. The listener may then relate the representation directly to the own articulator movements. Second language learners may also find it difficult to perceive or produce phonetic contrasts that do not exist in the mother tongue. Both groups could hence benefit from an augmented reality display of the face that shows the position and movement of intra-oral articulators together with the speech signal. Such an application has been developed in a joint showcase by KTH and LORIA, Nancy, France, within the European Network of Excellence MUSCLE.

Since we are normally unaccustomed to seeing the movements of the intra-oral articulators, it remains an open question if such information may be efficiently employed by listeners. Tarabalka et al. [4] showed that subjects who tried to identify the consonant in VCV words in noise did indeed perform better when they were presented with a midsagittal animation of the tongue movements together with the noisy speech signal, than when the speech signal was presented alone. They however also found that the identification was even better when the subjects saw the face with skin texture, rather than the tongue animations. This illustrates that, at least in their experiment, subjects found it easier to employ less detailed, but familiar visual information. Similarly, Graunwinkel et al. [5] concluded that the additional information provided by animations of the tongue, jaw and velum were not, in themselves, sufficient to improve the consonant identification scores in a similar task with VCV words in noise. Interestingly, both studies [4, 5] on the other hand found that subjects who received explicit or implicit training on how to interpret tongue movements performed significantly better than subjects who did not. Explicit training was used in [5], where one group of subjects was shown a video presentation explaining differences in place and manner of articulation for different consonants. Implicit training occurred for one group in [4], since they were first presented with test stimuli in clear acoustic conditions. These two recent studies therefore indicate that the display of tongue movements may improve consonant perception, if the subjects are first allowed to grow accustomed to the new type of visual information. The implications of the two studies for general speech perception are nevertheless limited, since only forced-choice identification of consonants was tested. If the articulatory display is to be used as an alternative to cued speech, we need to investigate if the intra-oral visualization can improve perception of a more complex content. In this study we therefore test a talking head with intra-oral animations as a support for word recognition in sentences.

2. The augmented visualization head

The MUSCLE visualization display consists of a double view of the face, from the front and the side, with transparent cheek, as shown in Fig. 1. The face [6], jaw and tongue [7] models are based on static 3D-wireframe meshes that are deformed by parametric weighted transformations [6]. The parameters for the face are jaw opening, shift, and thrust, lip rounding, upper lip raise, lower lip depression, upper lip retraction and lower lip retraction. For the tongue, they are dorsum raise, body raise, tip raise, tip advance and width. The tongue model is based on a statistical analysis of Magnetic Resonance Images of a Swedish subject producing vowels and consonants in three symmetric vowel-consonant-vowel (VCV) contexts [7]. Synthesized face and tongue movements can be created from a string of phonetic characters as input, using a rule-based audiovisual synthesizer. Interpolated parameter trajectories are created from the phoneme strings, taking visual coarticulation into account [8].



Name user08

skjortan fastnade på en spik

Figure 1. *Visual interface in the speech perception test. Three conditions were tested: (AO) Neither of the two faces shown; (AF) Front view of the face shown; (AFT) both front and side view shown.*

For the tongue movements, the coarticulation and timing of is modelled on Electromagnetic Articulography data [7].

The MUSCLE visualization head instead takes an acoustic utterance as input, generates the corresponding articulatory movements in the synthetic face and presents the animations synchronized with the acoustic signal. The articulatory movements are synthesized based on phoneme recognition of the spoken utterance.

In the perception test reported here, speech recognition is not used. Instead, the input to the visual speech synthesis is force-aligned label files of the pre-recorded utterances.

3. Stimuli and Subjects

The stimuli consisted of 60 short Swedish sentences spoken by a male Swedish speaker. The sentences have a simple structure (subject, predicate, object) and “everyday content”, such as “Kappan hänger i garderoben” (The coat hangs in the wardrobe) or “Laget förlorade matchen” (The team lost the game). These sentences are part of a set of 270 sentences designed for audiovisual speech perception tests by Öhngren, based on MacLeod and Summerfield [9]. The sentences are normally articulated and the speech rate was kept constant during the recording of the database by prompting the speaker with text-to-speech synthesis set to normal speed.

The sentences were presented in three different conditions: Acoustic Only (AO), Audiovisual with Face (AF), Audiovisual with Face and Tongue (AFT). For all conditions the acoustic signal was degraded and the audio-only condition provides a baseline intelligibility level. Two levels of audio degradation were used, to study the benefit of the visual information at two different simulated levels of hearing loss.

A noise-excited channel vocoder with 2 or 3 frequency channels was used to reduce the spectral details and create an amplitude modulated and bandpass filtered speech signal consisting of multiple contiguous channels of white noise over a specified frequency range [10]. The test was set up so that the 30 first stimuli were presented with three frequency channels and the 30 last with only two. The difficulty of the task was hence increased halfway into the test for all subjects. For the AF presentation a frontal view of the synthetic face is displayed and the AFT presentation in addition shows a side view, where intra-oral articulators have been made visible by making parts of the skin transparent (c.f. Fig. 1).

18 normal-hearing subjects participated in the experiment. All were current or former university students and staff. They were divided into three groups, as outlined in Table 1. The only difference between the three groups was that the sentences were presented in different conditions to different groups, so that every sentence was presented in all three conditions. The sentence order was random, but the same for all subjects. Since the degradation of the acoustic signal changed halfway into the test, the groups differed slightly regarding how many sentences in the two acoustic conditions that they were presented with, as indicated in Table 1.

Table 1. *Division of subjects and sentences. The distribution between 3- and 2-channel vocoded sentences is indicated for each set.*

	Set 1:	Set 2:	Set 3:
Group 1	AO	AF	AFT
Group 2	AFT	AO	AF
Group 3	AF	AFT	AO

Set 1: Sentences # 10 *vs.* 10	5, 6, 7, 10, 11, 12, 14, 19, 24, 27, 31, 32, 33, 37, 41, 46, 49, 51, 53, 60
Set 2: Sentences # 11 *vs.* 9	1, 2, 4, 8, 15, 16, 17, 18, 20, 23, 29, 34, 39, 40, 42, 44, 45, 48, 50, 57
Set 3: Sentences # 9 *vs.* 11	3, 9, 13, 21, 22, 25, 26, 28, 30, 35, 36, 38, 43, 47, 52, 54, 55, 56, 58, 59

4. Experimental set-up

The graphical interface for the perception test, shown in Fig. 1, consisted of an upper display with the animations of the face in front view (the AF case) or the face in a front view and a side view of the tongue and jaw (the AFT case).

The lower part of the interface was for the test subjects’ use, to start the test, repeat stimuli and type in their answers. The task was to identify as many words as possible in each sentence. The subjects were allowed to repeat the stimuli as many times as they wished before giving their answer. Repetitions were allowed since the sentence test material is quite complex and involves rapid tongue movements. Some subjects looking at VCV words in [4] further reported that it was difficult to simultaneously watch the movements of the lips and the tongue in *one* side view. In [5], each stimulus was presented three times before the subject should answer. Allowing repetitions made it possible for the subjects to focus on different audiovisual features for subsequent repetitions.

The acoustic signal was presented over headphones and the graphical interface was displayed on a 15” laptop computer screen. The perception experiment started with a familiarization set of sentences in AFT condition. The subjects were instructed to prepare for the test by listening to a set of five vocoded and five clear sentences accompanied by the double view of the synthetic face. Each familiarization sentence could be repeated as many times as the subject wanted. The correct answer could then be displayed upon request from the subject in the familiarization phase (no feedback was given on the subjects’ answers during the test). When the subjects felt prepared for the actual test, they started it themselves. The entire experiment, including familiarization and test, lasted 30-40 minutes.

5. Data analysis

The subjects’ written replies were saved in XML format and then analyzed manually. For each stimuli sentence, the presentation condition (AO, AF, AFT), the number of times the stimuli was played and the subject’s answer was stored together with the correct sentence text. The word accuracy was then counted disregarding morphologic errors.

The analysis focused on relating the word accuracy scores especially to the factors: presentation condition (Did visual cues improve the word recognition score?), stimuli sentence (Were some sentences better recognized in one condition? If so, why?), acoustic degradation (Was there any difference in visual contribution between the two levels of acoustic information?) and number of repetitions (Did additional visual information require more repetitions?).

6. Results

Fig. 2 shows the overall scores for the three different conditions, averaged over subjects. The results for the two audiovisual conditions were better than the acoustic only for both levels of audio degradation. A two-sided t-test showed that the differences were significant at a level of p<0.05 for three channels and p<0.0005 for two channels. The performance on the two audiovisual conditions was almost identical: AF (standard deviation SD=19 for three channels 3C, SD=20 for two channels 2C) and AFT (3C: SD=15, 2C: SD=14). Overall, the augmented reality display of the tongue movements did hence not improve the performance. Fig. 3 however shows that the performance differed substantially between the groups, with higher accuracy in AFT condition than in AF for the three channels signal for groups 1 and 2, but lower for group 3. There were further qualitative differences between the two- and three-channel conditions. This suggests that the phonetic content or semantic complexity varied between the different sentences. The mean performance on each sentence in the different conditions was therefore analyzed. Fig. 4 shows the difference in word accuracy rate between the two audiovisual conditions and the acoustic only (bars in the positive range hence indicating a better performance in AF and AFT compared to AO). From Fig. 4, one can identify the sentences for which AFT was much better than AF (sentences 9, 10, 17, 21, 22, 28, 30, 35, 37, 42, 56) and vice versa (1-3, 6, 12, 27, 33, 52, 57). A first observation concerning this comparison of the two audiovisual conditions is that of the first eight sentences, seven were more intelligible in the AF than in the AFT condition. This suggests that the subjects were still unable to use the additional information from the AFT display, despite the familiarization set, and were only confused by the tongue animations. The more intuitive AF view did on the other hand allow the subjects to perform better through lip reading.

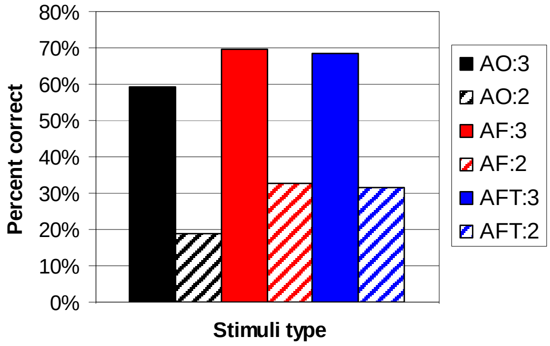


Figure 2. *Mean word identification scores for the three conditions AO: Acoustic Only, AF: Audiovisual with Face, AFT: Audiovisual with Face and Tongue. Filled bars refer to three-channel vocoded speech, striped to two-channels.*

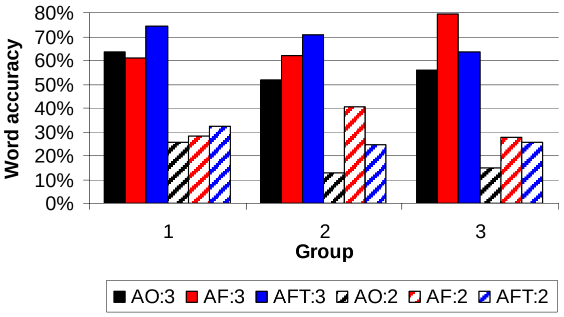


Figure 3. *Mean sentence intelligibility per group.*

For all but two (17 and 42) of the sentences that were more intelligible with the AFT display than the AF, the difference in word accuracy score is between groups 1 and 2. Since the overall AFT score for these two groups were better than the AF score, the differences may be attributed to the phoneme sequences in the test sentences.

An analysis of the sentences that were better perceived in AFT than in AF condition indicates that subjects found additional information in AFT mainly for palatals [k, g], the liquid [l] and the rhotic [r]. In particular for consonant clusters combining palatal plosives and liquids, [kl, rk], but also other clusters with [l, r], such as [dr, tr, lj, ml, pl, pt], were better perceived with animations of the tongue. The effect was not universal, for all occurrences of the clusters or all subjects. The results nevertheless suggest that subjects were able to extract information from the animation of the raising of the tongue tip (for [r, l]) or tongue dorsum (for [k, g]), that may be difficult to perceive from a front face view.

The sentences that were better perceived in the AF condition contained more bilabials and labiodentals than the average sentences, indicating that the AFT improvement comes at a price: Subjects may miss information that is clearly visible in the AF view when concentrating on the tongue.

The number of repetitions used in different display conditions is shown for the three channel vocoded speech in Fig. 5. The graph shows a bimodal distribution of the number of repetitions. The subjects were either certain about their perception after 1-3 repetitions, or they used many (more than six in 30% of the stimuli) to try to decode difficult sentences. This division into two different behaviors was clearer for the two audiovisual conditions than the audio-only, and even clearer for the AFT case than the AF. The addition of the augmented reality side-view of the face hence made the subjects more confident about their perception. The same observation can be made for the stimuli with two voccoded channels: the audiovisual conditions either required fewer repetitions or led the subjects to repeat the stimuli many more times than in the acoustic only presentation. Thus, over 50% of the more severely acoustically degraded sentences were played more than six times when visual information was given (52% for AF, 57% for AFT). For the acoustic only condition, the share of many repetitions was the same with both degrees of degradation (AO 32%). It should be noted that the word accuracy decreases for each additional repetition (from around 90% for one or two repetitions to just over 30% for more than six repetitions for all conditions in the three channel case), and subjects hence gained little by repeating the stimuli many times. The AFT condition was slightly different from the two others in that the recognition performance was best when the stimulus was played twice (88% words correct, compared to 78% for one repetition). It

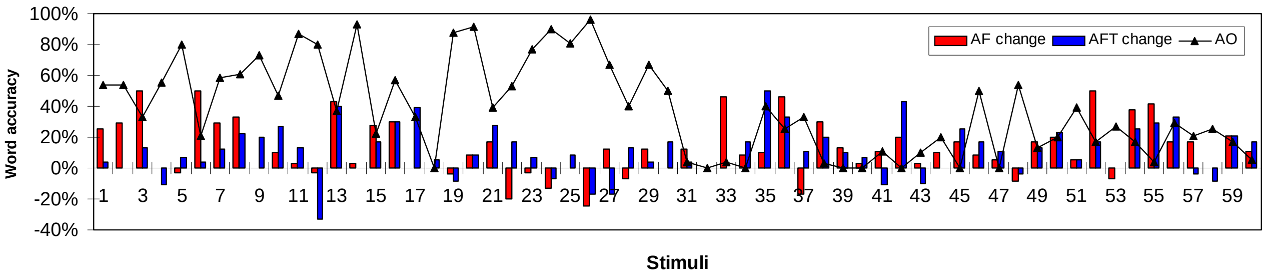


Figure 4. *The mean contribution for each sentence of the synthetic face in the two audiovisual conditions in relation to the performance on the audio-only condition (line shows mean AO accuracy). Stimuli 1-30 has vocoded speech with 3 channels, stimuli 31-60 with 2.*

appears that the additional repetition allowed users to take information from both face views into account.

7. Discussion and Conclusions

This preliminary test has shown that subjects overall get no additional support in sentence perception from an augmented reality view of the face, displaying tongue movements from the side, compared to a front view of the face. This is not a surprising finding, since the animated tongue movements are unfamiliar, whereas speech reading of a synthetic face can build on human face-to-face communication. It is also in line with the findings in [4]. Some test sentences, containing palatal plosives and/or liquids/rhotics, were nevertheless better perceived when the subjects were presented with tongue animations. It thus appears that subjects are in fact able to learn to extract some information about phonemes from the intra-oral articulation, even in a more complex speech material consisting of sentences. Inter-subject variability was very high. Some subjects did clearly benefit from the AFT view, with up to 30% better word recognition compared to the AO or AFT cases. Another subject from the same group however scored 48% lower in AFT than in AF for the same sentences (two-channels). Some subjects are hence very receptive to the AFT view, but due to the rapidity of the tongue movements it seems unrealistic that articulatory information can be used as an alternative to cued speech for real-time speech perception, without large amounts of training. It is, on the other hand, easier to envisage that intra-oral articulation displays can be beneficial in computer-assisted pronunciation and perception training applications [11], where the user can repeat the animations the desired number of times or even play them in slow motion. In such an application, the additional articulatory information conveyed by the intra-oral animation may support the user in establishing the articularoty-acoustic relationship for the foreign phonemes.

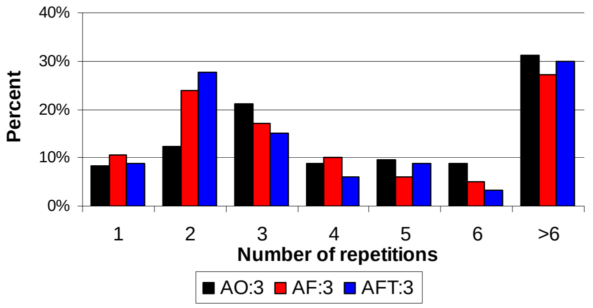


Figure 5. Mean share of different numbers of repetitions for the different condition.

8. Acknowledgements

The visual display used in the perception test was partially developed in the Network of Excellence MUSCLE (Multimedia Understanding through Semantics, Computation and Learning), funded by the European Commission.

9. References

[1] Sumby, W.H. and Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. Journal of the Acoustical Society of America, 26, 212-215.

[2] 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.

[3] Cornett, O. & Daisey, M. E. (1992). The Cued Speech Resource Book for Parents of Deaf Children. National Cued Speech Association.

[4] Tarabalka, Y., Badin, P., Elisei, F. and Bailly, G. (2007). Can you ”read tongue movements”? Evaluation of the contribution of tongue display to speech understanding. Proceedings of ASSISTH2007, 187-193.

[5] Grauwinkel, K., Dewitt, B. and Fagel, S. (2007). Visual Information and Redundancy Conveyed by Internal Articulator Dynamics in Synthetic Audiovisual Speech. Proceedings of Interspeech 2007, 706-709.

[6] Beskow, J. (1997). Animation of Talking Agents, Proceedings of AVSP'97, 149-152.

[7] Engwall, O. (2003). Combining MRI, EMA & EPG in a three-dimensional tongue model, Speech Communication, vol. 41/2-3, 303-329.

[8] Cohen, M. and Massaro, D. (1993). Modeling coarticulation in synthetic visual speech". In D. Thalmann N. Magnenat-Thalmann, (eds.), Computer Animation '93. Springer-Verlag.

[9] MacLeod, A., and Summerfield, Q. “A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise. Rationale, evaluation and recommendations for use", *British Journal of Audiology*, 24:29-43, 1990.

[10] Siciliano, C., Williams, G., Beskow, J., & Faulkner, A. (2003). Evaluation of a Multilingual Synthetic Talking Face as a communication Aid for the Hearing Impaired. In Proc of Intl Conf of Phonetic Sciences, pp. 131-134

[11] Engwall, O., Bälter, O., Öster, A-M., and Kjellström, H. (2006). Designing the user interface of the computer-based speech training system ARTUR based on early user tests. Journal of Behavioural and Information Technology, 25(4), 353-365.