Looking at tongues – can it help in speech perception?

Preben Wik, Olov Engwall
Centre for Speech technology, School of Computer Science and Communication, KTH, Sweden

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.

Introduction

It is well established that visual information support speech perception, especially if the acoustic signal is degraded (Sumby and Pollack 1954). Not only hearing-impaired listeners but also normal-hearing listeners benefit from information given by the face, and it has been shown, in e.g. Agelfors et al (1998), 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. Would it be beneficial to supplement the acoustic signal and speech reading of the face with additional visualization of tongue movements? An application has been developed in a joint showcase by KTH and LORIA, Nancy, France, within the European Network of Excellence MUSCLE, to investigate the potential benefit of such an augmented reality display with two groups in mind.

1) A community of hearing-impaired persons that rely on cued speech, (where additional phonetic information is conveyed with hand sign gestures) (Cornett and Daisey1992).

2) Second language learners that may find it difficult to perceive or produce phonetic contrasts that do not exist in the mother tongue.

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.

Two recent experiments (Tarabalka et al. 2007; Graunwinkel et al. 2007) investigated if consonant identification in CVC words could be enhanced by animations of tongue movements when the speech signal was noisy. The studies showed that the display of tongue movements did improve consonant perception, but only if the subjects were 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 a speech perception support, 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.

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, jaw and tongue models are based on 3D-wireframe meshes that are deformed by parametric weighted transformations (Beskow 1997). 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 (Engwall 2003). 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 (Cohen and Massaro 1993). For the tongue movements, the coarticulation and timing is modelled on Electromagnetic Articulography data (Engwall 2003).

The MUSCLE visualization head takes an acoustic utterance as input, generates the corre-

sponding 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.

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

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 (1990).

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 (Siciliano et al 2003). 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.

18 normal-hearing subjects participated in the experiment. All were current or former university students and staff. They were divided into three groups, where the only difference between the three groups was that the sentences were presented in different conditions to different groups. This was made so that every sentence was presented in all three conditions. The sentence order was random, but the same for all subjects.

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 and a lower part where the test subjects could 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.

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.

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?).

Results

Fig. 2 shows the overall scores for the three different conditions, averaged over subjects in the tree different groups. 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 not improve the performance. Fig. 2 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. 3 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. 3, 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).

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 consonant clusters with 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.

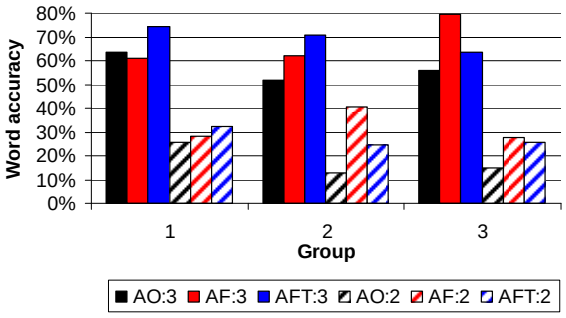

Figure 2. Mean word identification scores per group, and for the three conditions AO: Acoustic Only, AF: Audiovisual with Face, AFT: Audiovisual with Face and Tongue. The filled bars refer to three-channel vocoded speech, and the striped to two-channels (more acoustically degraded).

Discussion and Conclusions

Overall the tongue movement display did not give any additional support in sentence perception, 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. Some test sentences that contained phonemes with attributes that are invisible in the face, such as palatal plosives and/or liquids/rhotics, were better perceived when the subjects were presented with tongue animations. It thus appears that

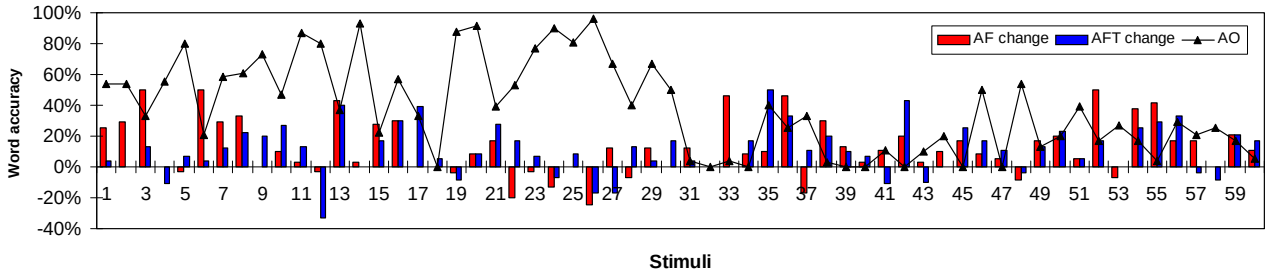

Figure 3. The mean contribution for each sentence of the synthetic face in the two audiovisual conditions in relation to the performance on the audio-only condition (line shows mean AO accuracy). Stimuli 1-30 has vocoded speech with 3 channels, stimuli 31-60 with 2.

subjects are in fact able to extract some information about phonemes from the intra-oral articulation, even in a more complex speech material consisting of sentences. Some subjects did however clearly benefit from the AFT view, with the best subjects having 30% better word recognition compared to the AO or AF cases. Inter-subject variability was however very high, and another subject from the same group scored 48% lower in AFT than in AF for the same sentences (two-channels).

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, due to the rapidity of the tongue movements. It is easier to envisage that intra-oral articulation displays can be beneficial in computer-assisted pronunciation and perception training applications, (Engwall et al 2006) 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.

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. The research was also supported by the Graduate School of Language Technology (GSLT).

References

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., & Öhman, T. (1998). Synthetic faces as a lipreading support. Proceedings of ICSLP.

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

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

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

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

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.

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.

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.

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

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

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.