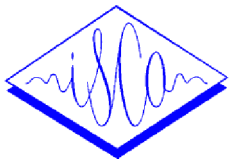
ISCA Archive
http://www.isca-speech.org/archive



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

Effects of Smiled Speech on Lips, Larynx and Acoustics

Sascha Fagel

Berlin Institute of Technology,
sascha.fagel@tu-berlin.de

Abstract

The present paper reports on results of a study investigating changes of lip features, larynx position and acoustics caused by smiling while speaking. 20 triplets of words containing one of the vowels /a:/, /i:/, /u:/ were spoken and audiovisually recorded. Lip features were extracted manually as well as using a 3D motion capture technique, formants were measured in the acoustic signal, and the vertical larynx position was determined where visible. Results show that during production of /u:/ F1 and F2 are not significantly affected despite of changes of lip features while F3 is increased. For /a:/ F1 and F3 are unchanged where for /i:/ only F3 is not affected. Furthermore, while the effect of smiling on the outer lip features is comparable between vowels, inner lip features are differently affected for different vowels. These differences in the impact on /a:/, /i:/ and /u:/ suggest that the effect of smiling on vowel production is vowel dependent.

Index Terms: smiled speech, motion capture, vowel production

1 Overview

Many aspects of speech are affected when the speaker is smiling compared a neutral expression. The changes in speech production lead to observable differences in the acoustic and optic signals as well as to an altered perception. This has strong implications on speech recognition and speech synthesis with respect to both acoustic and visual modalities. The following section describes previous findings in production, acoustics and perception of speech that are relevant to smiled speech. The remaining sections of the paper describe method and results of a study on the effect of smiling on acoustic and facial properties of speech. Previous and current findings are integrated in the conclusions.

2 Previous findings

2.1 Effects of Smiling on Speech Production

Fant [1] mentioned that lip protrusion in rounded vowels lengthens the vocal tract and this way lowers formant frequencies. The reverse relationship was described by Shor [2] in smiled speech: the mouth widens and the lips retract resulting in a shortened vocal tract. Hence, smiling during speech constitutes a conflicting demand on the lip shape, at least in rounded vowels. Humans need to have a strategy to maintain the acoustic properties of the speech signal so that it can still be decoded by a listener.

Riordan [3] found that larynx lowering can compensate for reduced lip rounding due to smiling during speech. However, no compensatory larynx lowering was found in unrounded vowels. Another – presumably the main function – of vertical larynx position is F0 control (e.g. Honda et al. [4]). Savariaux et al. [5] found in a lip tube experiment that the tongue position can compensate for reduced lip rounding. All three articulatory parameters – vertical larynx position, tongue position, and lip spreading – contribute to the effective length of the vocal tract to different degrees. In a study of radiographic and labial films Dusan [6] investigated the vocal tract length during the production of consonants and vowels and found correlations to lip protrusion with $r_{consonants}$=0.72 and r_{vowels}= 0.77, to the position of tongue dorsum with $r_{consonants}$=0.63 and r_{vowels}=0.74, and with the lowest correlation among these three parameters to vertical larynx position with $r_{consonants}$=0.63 and r_{vowels}=0.65.

2.2 Acoustic Consequences and Effects of Smiling on Speech Perception

The properties of speech production associated with smiled speech and possible compensations and their effect on format frequencies can be analyzed by a simulation using an articulatory speech synthesizer (preferably the one from Birkholz [7]). Lip protrusion lowers F1 and F2 and only to a marginal degree F3. The reduction of the vertical lip opening mainly decreases F2 but also lowers F1 and F3. Moving the tongue dorsum backwards lowers F2. Larynx lowering has a main effect in lowering F3 and also lowers F2. In a simple model of speech production (Patterson et al. [8]) the cavity in front of the constriction (closer to the lips) mainly accounts for the second formant, the first and third formant are mainly produced by the back cavity, i.e. behind the constriction (closer to the larynx). This simplification describes the formant frequencies better the narrower the constriction between front and back cavities. Furthermore, tongue and larynx are connected with one another through the hyoid bone and hence do not move completely independently.

Not surprisingly, smiling in speech can be identified auditorily (e.g. Tartter [9]); an increased amplitude of speech leads to robust identification of that speech as smiled – although speakers do not produce speech with significantly increased amplitude when smiling. Where Tartter [9] found an increased F0 in smiled speech compared to non-smiled speech, in a repetition of that study Tartter & Brown [10] did not find such an effect. However, it was shown that smiling while speaking leads to increased formant frequencies. The effect on segmental duration was not significant.

Lasarcyk & Trouvain [11] varied several speech production parameters in a study using an articulatory speech synthesizer [7]. The results showed that synthetic speech with raised larynx was identified as more smiled – even if the spreading of the lips was kept constant. In a study by Drahota et al. [12] audio stimuli with higher F0 were perceived as more smiled even though this effect did not occur significantly in speech produced while smiling (whether or not a stimulus was produced with a smile was identified visually). Analogously, stimuli with higher intensity were

perceived as more smiled although this parameter was not significantly changed by smiling in the speech production.

As a consequence of the abovementioned perception studies the question arises what defines smiled speech as smiled. In a study on auditory-visual perception of expressive speech [13] Fagel cross-combined the audio and video tracks of utterances that were perceived as *happy, angry, sad,* and *content* when presented in a single modality (audio alone or video alone). However, a *content* voice combined with an *angry* face was perceived as *sad*, and an *angry* voice combined with a *content* face was perceived as *happy*. A possible explanation is that the valence (i.e. positive/negative evaluation) is transmitted predominantly by the face and the activation (i.e. arousal/relaxation) is transmitted predominantly by the speech audio. Hence, cross-modal effects can be expected in the perception of smiling in speech, too.

Edmond et al. [14] listed further possible influences on the perception of smiling – lexical, cultural, gender, and speaker effects – which have yet to be investigated.

3 Experimental Setup

3.1 Corpus and Material

A list of 20 word triplets with each word containing one of the vowels /a:/, /i:/, /u:/ was compiled in analogy to minimal pairs. The word triplets were selected among a longer list of possible triplets in order to contain the vowels under investigation in a wide variety of consonantal contexts with respect to different places and manners of articulation. The words are

1.	Maß	mies	Mus
2.	Basen	Biesen	Busen
3.	Paar	Pier	pur
4.	fahren	vieren	Fuhren
5.	Saale	Siele	Suhle
6.	Saat	sieht	Sud
7.	nah	nie	Nu
8.	da	die	du
9.	dar	dir	Dur
10.	Stahl	stiehl	Stuhl
11.	lad	Lied	lud
12.	blasen	bliesen	Blusen
13.	Schale	schiele	Schule
14.	Schar	schier	Schur
15.	Schaf	schief	schuf
16.	Gras	Grieß	Gruß
17.	graben	Grieben	Gruben
18.	brat	briet	Brut
19.	Hafen	hieven	Hufen
20.	haben	hieben	huben

The 60 words were spoken by a male speaker with the instruction to speak them once with neutral expression and once while smiling. In case of the smiled speech the speaker tried to keep the smile constant over each entire single word utterance. Four cameras (DragonflyExpress from Point Grey Research) recorded the face of the speaker synchronously at 60 frames per second: three cameras with left center and right view of the face and one camera recording the larynx region against black background. 44 colored markers were attached to the speaker's face. The audio speech was recorded with a head mounted microphone (AKG C 420L + B29L) at a distance of 15cm to the lips. The setup is shown in Figure 1, the four camera views are shown in Figure 2.

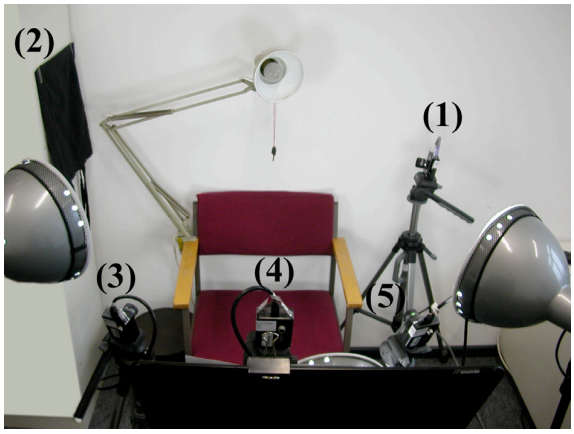


Figure 1: Experimental setup with larynx camera (1) against black background (2), left (3), center (4) and right (5) camera.

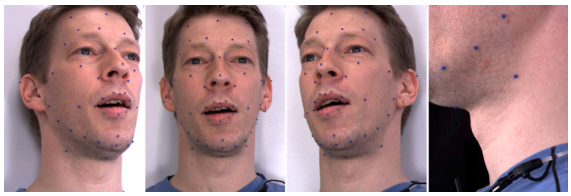


Figure 2: The four camera views on the subject.

3.2 Measurements

12 measures were derived from the recordings.

3.2.1 3D Data

The audio track of the recordings were annotated on phone level in order to obtain the center of the realization of the vowels under investigation. The markers that were glued on the face of the speaker were once registered manually using the web-based CLIC'N'TRAK [15] 3D motion capture software. This is done by clicking once each marker in one reference image from each of the three camera views (left, center, and right). The reference between the three views on one marker were defined by clicking all markers in the same order.

The 2D positions of the markers in the camera images were obtained by automatic marker tracking in the image sequences with manual corrections where necessary. The system provides the centers of the near-spherical markers on pixel level. The image resolution is 640x480 pixels at a captured face area of about 32x24cm (face cameras) in the present study. Hence, the technical accuracy of the system is 0.5mm (the effective accuracy was not determined). 3D marker positions were calculated by triangulation of the three 2D coordinates of each marker using projection matrices of the cameras (that were obtained by camera calibration on a known calibration object). The following measures were derived from the 3D positions of the markers on the upper and lower lip centers, lip corners and check bones.

- Lip spreading: 3D Euclidean distance between lip corners in mm
- Vertical lip opening: 3D Euclidean distance between upper and lower lip centers in mm
- Absolute lip protrusion: 3D Euclidean distance between the midpoints of the direct connection of the upper and lower lip centers and the direct connection of the cheek bones in mm

- Relative lip protrusion: 3D Euclidean distance between the midpoints of the direct connection of the upper and lower lip centers and the direct connection of the lip corners in mm

3.2.2 2D Data

As the marker data only provides information about the outer lip contour, the inner lip contour was manually marked by the use of an image editor. Three measures were calculated from this inner lip contour.

- Inner lip opening width: 2D Euclidean distance between left and right inner lip corners in pixels
- Inner lip opening height: average lip opening height calculated on the inner lip contour normalized to horizontal connection line between inner lip corners in pixels
- Inner lip area: amount of pixels surrounded by the inner lip contour in square pixels

The images of the centers of vowel realization of the fourth camera that recorded the larynx region were marked manually by the use of the CLIC'N'TRAK software. The derived measure was the

- Vertical larynx position: 2D Euclidean distance between Adam's apple and a skin point on the bottom of the neck in pixels

3.2.3 Acoustic Measures

The fundamental frequency (F0) and the first three formants (F1 – F3) were extracted automatically with praat [16] at the centers of vowel realization. In some /u:/ realizations only one formant at F1/F2 was extracted, so the data were manually checked and corrected where necessary.

4 Results

Table 1 shows the mean values and the significance levels of the measurements in non-smiled versus smiled speech for all three vowels. All mean values whose differences between non-smiled and smiled speech were non-significant are shown in bold face. Vertical larynx position could not be measured for the vowels /a:/ and /i:/ because the larynx disappeared upwards out of the observable range. For /u:/ the larynx was raised in smiled speech compared to non-smiled speech.

4.1 Outer Lip Measures

Lip spreading increased and vertical lip opening decreased from non-smiled to smiled speech for all three vowels with the effect on vertical lip opening being the smallest for /i:/. The between-vowel differences of all mean values are significant with the differences in lip spreading becoming small in smiled speech (56.6mm ± 1.3mm) but still significant ($p<.05$; not listed in the table).

4.2 Lip Protrusion

Relative lip protrusion increased and absolute lip protrusion decreased from non-smiled to smiled speech for all three vowels at comparable degrees. The absolute lip protrusion decreased more if the vowel itself was more protruded. The between-vowel differences in non-smiled speech are significant ($p<.001$; not listed in the table). The between-vowel differences in smiled speech are significant at a lower levels ($p<.05$ to $p<.001$; not listed in the table) or non-significant for the difference between /a:/ and /u:/, respectively.

4.3 Inner Lip Measures

The inner lip width increased from non-smiled to smiled speech for all vowels with the highest percentual increase for /u:/. The inner lip opening height is decreased for /a:/ and /u:/ but not for /i:/ what is inline with the small change of outer vertical lip opening for /i:/. As a result of these two effects the inner lip area is not significantly increased for /a:/ (increased width, decreased height), is significantly increased for /i:/ (increased width, non-decreased height), and is largely increased for /u:/ (largely increased width not compensated by decreased height).

4.4 F0 and Formants

F0 increased from non-smiled to smiled speech. Vowels can only be limitedly compared with one another as the order of vowels in a three-words sequence (e.g. "lad – Lied – lud") was not randomized and hence a prosodic effect on "utterance" level cannot be excluded. The increase was least for /a:/, higher for /i:/ and most for /u:/.

Table 1: Mean values and the significance levels (ANOVA) of the measurements in non-smiled versus smiled speech for the vowels /a:/, /i:/ and /u:/. Non-significant differences between non-smiled and smiled speech are shown in bold face.

	/a:/			/i:/			/u:/		
measure	sig.	non-smiled	smiled	sig.	non-smiled	smiled	sig.	non-smiled	smiled
lip spreading (mm)	.000	45.1	56.6	.000	49.9	57.8	.000	37.5	55.3
vertical lip opening (mm)	.000	46.9	37.3	.002	29.9	28.0	.000	34.1	23.8
absolute lip protrusion (mm)	.000	96.6	93.0	.000	93.1	91.1	.000	98.1	93.2
relative lip protrusion (mm)	.000	15.6	21.3	.000	16.9	21.9	.000	14.1	20.9
inner lip width (pixels)	.000	87.78	119.9	.000	79.3	114.0	.000	16.6	48.4
inner lip height (pixels)	.000	41.4	30.7	.892	**17.8**	**18.0**	.001	6.0	4.1
inner lip area (sq. pixels)	.607	**3742.9**	**3807.1**	.000	1495.8	2142.7	.000	118.6	250.3
F0 (Hz)	.000	104.0	149.0	.000	124.5	192.2	.000	124.8	213.3
F1 (Hz)	.852	**746.2**	**744.4**	.000	262.8	301.0	.248	**290.4**	**305.6**
F2 (Hz)	.000	1136.6	1245.6	.001	2004.1	2099.6	.736	**767.6**	**777.4**
F3 (Hz)	.874	**2273.9**	**2293.6**	.083	**2829.8**	**2886.8**	.001	2191.3	2299.7
vertical larynx position	--	--	--	--	--	--	.000	43.5	65.3

For /a:/ and /i:/ F3 does not raise significantly from non-smiled to smiled speech. For /a:/ also F1 is not increased. /u:/ shows a different formant pattern: F1 and F2 are not significantly increased but F3 is increased.

5 Conclusions

As expected smiling results in stretched lips with smaller differences between vowels in smiled compared to non-smiled speech. At the same time the vertical lip opening is reduced. Inner lip measures do not show such a consistent picture: smiling affects inner lip height (/a:/), inner lip area (/i:/), or both (/u:/). The decreasing level of significance of lip spreading and protrusion of between-vowel differences in smiled speech compared to non-smiled speech suggests that the lip spreading reaches a level of saturation. Furthermore, the increase of relative protrusion and the decrease of absolute protrusion indicates that the lip centers reach this level of saturation earlier when they get deformed by contact to the incisors while the lip corners still can move further backwards. It is assumed that the outer lip contour maintains consistent visibility of smiling where the inner lip contour through an adapted lip shape) helps to maintain the acoustics in order to preserve the identification of vowel qualities. A perceptual evaluation study is planned to further investigate this conclusion.

Assuming that smiling has a major effect on the front cavity, the formant patterns of /a:/ and /i:/ can mostly be explained by the fact that the cavity in front of the constriction mainly accounts for the second formant while the first and third formant are mainly produced by the back cavity. However, the increase in F1 for /i:/ is not supported by this explanation. The increased F0 is assumingly accompanied by a raised larynx (but could not be measured in the present study) which can partly account for the increased F1. It is remarkable that in the case of the rounded vowel /u:/ the vowel quality by F1 and F2 is not significantly affected by smiling where F3 is significantly increased. As the acoustically relevant inner lip area is increased, this indicates a presence of a compensation. This supports the previous finding that compensation occurs predominantly in rounded vowels. Whether this compensation is realized by tongue position adjustment cannot be answered in the present study due to the lack of tongue data.

Smiling accesses in parts physiological mechanisms that are also involved in the production of speech which results in conflicting demand in case of rounded vowels. The effect of smiling on the measures of vowel realization used in the present study is highly significant in most cases. However, the effect appears differently on different vowels. This suggests that smiling combines with articulatory configurations in a vowel-dependent manner. This has strong implications on the visual synthesis of smiled speech. Using a single (linear) parameter that changes a neutral face into a smiling face and superimpose articulatory movements that were measured on a neutral face will not produce satisfactory results. Even adjusting the lip spreading parameter in smiled speech will not suffice due to the vowel dependency of the effect of smiling. Analogously, a phoneme-dependent approach is necessary for audio synthesis of smiled speech (e.g. when applying a voice conversion technique). However, in order to determine in detail the nature of the dependency between vowel production and smiling a study of a more complete set of vowels is required.

6 Acknowledgements

Many thanks to the members of the EU COST Action 2102 on "Cross-Modal Analysis of Verbal and Non-verbal Communication" for their valuable comments, and to Jürgen Trouvain for the basic idea of the present study and for a number of interesting and relevant references. This work was supported by the German Research Foundation DFG (FA 795/4-1).

7 References

[1] G. Fant, *Acoustic Theory of Speech Production,* Mouton, The Hague, 1960.

[2] R.E. Shor, "The production and judgment of smile magnitude" in *Journal of General Psychology* 98: pp. 79-96, 1978.

[3] C.J. Riordan, "Control of vocal-tract length in speech", in *JASA* 62, pp. 998–1002. 1977.

[4] K. Honda, H. Hirai, S. Masaki, Y. Shimada, "Role of vertical larynx movement and cervical lordosis in F0 control", in *Language and Speech* 4, pp. 401-11, 1999.

[5] C. Savariaux, P. Perrier, J.P. Orliaguet, "Compensating for labial perturbation in a rounded vowel: an acoustic and articulatory study", in *Proceedings of EUROSPEECH*, pp. 89-92, 1993.

[6] P. Birkholz, *3D-ArtikulatorischeSprachsynthese*. PhD. dissertation, University of Rostock, 2005.

[7] S. Dusan, "Vocal Tract Length during Speech Production", in *Proceedings of INTERSPEECH*, 2007.

[8] R. Patterson , J. Monaghan, T. Walters, "A simple, formant-pattern model of speech communication", retrieved on April 26 2009 from http://www.acousticscale.org/wiki/index.php/A_simple,_formant-pattern_model_of_speech_communication

[9] V. C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech", in *Perception and Psychophysics* 27(1), pp. 24–27, 1980.

[10] V.C. Tartter, D. Braun, "Hearing smiles and frowns in normal and whisper registers", in *JASA* 96(4), pp. 2101-2107, 1994.

[11] E. Lasarcyk, J. Trouvain, "Spread Lips + Raised Larynx + Higher F0 = Smiled Speech? - An Articulatory Synthesis Approach", in *Proceedings of ISSP*, 2008.

[12] A. Drahota, A. Costall, V. Reddy, "The vocal communication of different kinds of smile", in *Speech Communication* 51(4), pp. 278–287, 2008.

[13] S. Fagel, "Emotional McGurk Effect", in *Proceedings of Speech Prosody*, 2006.

[14] C. Émond, J. Trouvain, L. Ménard, "Perception of smiled French speech by native vs. non-native listeners : a pilot study", in *Proceedings of the Interdisciplinary Workshop on Phonetics of Laughter*, pp. 27-30, 2007.

[15] S. Fagel, "CLIC'N'TRAK – A Web-based 3D Motion Capture System", [Computer program], retrieved April 26 2009 from http://avspeech.info/clicntrak/

[16] P. Boersma, D. Weenink, "Praat: doing phonetics by computer (Version 4.3.14)", [Computer program], retrieved May 26 2005 from http://www.praat.org/