Spread Lips + Raised Larynx + Higher F_0 = Smiled Speech? – An Articulatory Synthesis Approach

Eva Lasarcyk, Jürgen Trouvain
Saarland University
E-mail: (evaly|trouvain)@coli.uni-saarland.de

Abstract

We present an initial study on how to model smiled speech with an articulatory speech synthesizer, led by the research question as to what cues are responsible for the effect of an audible distinction of smiled vs. non-smiled speech. In a perception test, we explore the relative contributions of i) spreading of the lips, ii) raising of the larynx, and iii) raising of the fundamental frequency.

36 test subjects assessed isolated synthetic vowel stimuli of /aː, iː, yː, uː/ on a 5-point "smiley scale". Results indicate that F_0 is the main acoustic factor for perceiving smileyness. The other factors depend on the vowel quality, with best results for the unrounded vowels /iː/ and /aː/.

1 Introduction

This study examines whether it is possible to model smiled speech with an articulatory synthesizer. Several studies report that smiled speech can be distinguished auditorily from non-smiled speech [10, 11, 9, 8, 2]. Parameters which were found to be typical of smiled speech comprise raised F_0 and raised formant values [10, 11, 8]. Increased values for F1 and F2 can be explained with a shortening of the vocal tract that occurs when the lip corners are retracted for smiling. This "i-face" in Ohala's frequency code [6] – in contrast to the "o-face" – is also suggested as a typical setting for signaling smallness of the speaker [13] by increased formant frequencies (indicating a smaller vocal tract) and increased phonation rate (indicating smaller vocal folds).

Rather unexplored (but cf. [13]) is the possibility of a shortening of the vocal tract by raising the larynx as has been observed for varying the vocal tract length during vowel production [7]. This can have an effect of a) raising the formant values (cf. [5, 13]) and b) raising F_0 used in Asian languages as part of register [3].

The present study seeks to find the relative contributions of the three parameters lips, larynx, and F_0 possibly responsible for the perceptual effect of smileyness in speech. In a similar but not identical study, [13] showed that a manipulation of these three parameters could be used as cues for body size and anger-joy distinction. With human speakers, in contrast to an articulatory synthesis approach, several methodological problems would be expected: Speakers vary the intensity of smiling [9, 2] which is also observable in the degree of lip spreading [8]. In addition, the effects of felt and non-felt smiles on speech are still unclear [9, 2]. Also, measuring larynx height is not a straightforward endeavour.

2 Stimulus Material

We used the articulatory synthesis system by [1] to create stimuli for a perception test. The synthesizer is based on Magnetic Resonance Imaging (MRI) and x-ray data of one German speaking male subject [1]. The German vowels /iː, aː, uː, yː/ were generated as single vowel utterances with a duration of 560 ms. F_0 was set at a monotone 112 Hz to avoid interactions with specific intonation contours possibly expressing disgust, although monotonous F_0 usually shows a stronger tendency to be judged as "sad" or "bored".

For each "neutral" vowel the following three changes were applied:

1. Spreading of the lips (retracting the lips to the most extreme position possible)

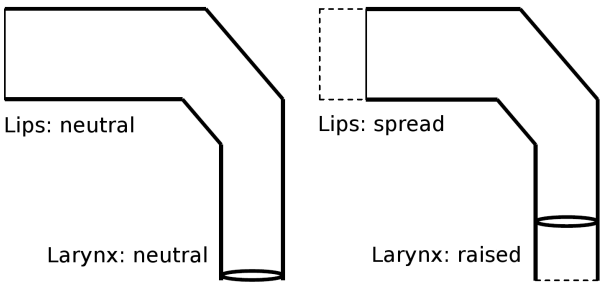

Figure 1: *Schematic illustration of vocal tract shortening: Neutral (left) vs. shortened shape (right).*

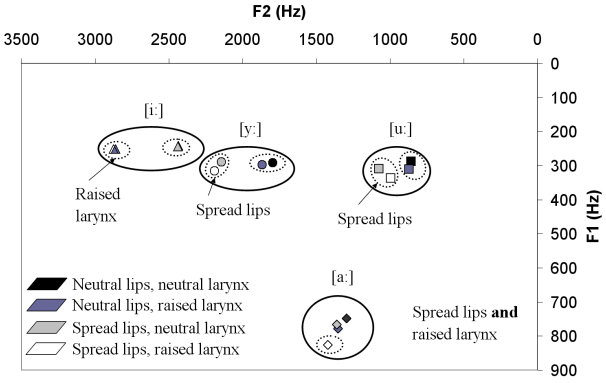

Figure 2: *Formant plot of all stimuli in high F_0 setting. Lip and Larynx parameters influence formants.*

2. Raising of the larynx (highest position possible combined with a slightly tenser voice quality, see also [5])

3. Raising F_0 higher than normal (increased by 2 semitones, in this case from 111 to 125 Hz)

The combination of all parameters yielded 32 different stimulus vowels (4x2x2x2=32) [12]. A schematic illustration of a neutral vocal tract vs. a vocal tract with spread lips and raised larynx configuration can be seen in Figure 1.

Since we were using an articulatory speech synthesizer in a novel way for smiled speech, information on its performance should be provided regarding the *vowel quality* in the different conditions.

Formant analysis yielded no relevant changes of formant frequency values when only F_0 was manipulated. As expected, however, lip spreading contributed to formant changes as well as larynx raising, as can be seen in Figure 2. For the two rounded vowels, spreading of the lips, i.e. shortening of the vocal tract at the front end, raises mainly F2. In /iː/ and less also /aː/, the raising of the larynx mainly contributes to a formant value increase. Here, the vocal tract could not be shortened any more at the lip end but the shortening took place at the larynx and it is reflected in the acoustics of the vowel.

Due to the applied manipulations one would expect possible confusions in the perception of vowel categories. In contrast to [13], we did *not* manipulate further parameters to preserve vowel quality. In a separate perception test, six phonetically trained participants marked which vowel *category* they *perceived* on the IPA vowel chart. The perception of vowel quality was most stable across all variants of /iː/. For the different /aː/ stimuli, the perception of a slight fronting resulted from spreading the lips, raising the larynx, or both. For /uː/, raising the larynx still suggested a stable categorization of /uː/. However, lip spreading, and both lip spreading and larynx raising lead to unstable perception results. The /yː/ variants revealed an unstable recognition even in the neutral form. Likewise, lip *or* larynx manipulations lead to unstable category perceptions; lip spreading *and* larynx raising stabilized the categorical perception to some extent but the sounds for /yː/ were assigned exclusively to *un*rounded vowels.

This confusion in the perception of the basic vowel category has to be taken into account in the interpretation of the results especially for the two rounded vowels. It was strongest for /uː/ and /yː/, which are both rounded, so that lip spreading for "smiling" might interfere with the basic vowel quality. Another influencing factor could be the vowel intrinsic larynx height, which is e.g. in the case of /uː/ rather low as compared to /aː/.

3 Perception Test

Subjects were asked to rate the perceived smileyness of the 32 vowel stimuli. The perception test was carried out as web-based experiment with invited subjects. The experiment started with an explicit warm-up phase: It did not serve as guidance on how to give answers, nor were any answers saved from this phase; it only familiarized the participants with the range of stimuli, layout, and technical process of the test set-up.

In the main experiment, the stimuli were presented in three rounds in randomized order. Using their home PC loudspeakers, 36 German speaking

subjects were asked to rate the stimuli on a five-point scale: "1" representing a vowel produced with "mouth corners pulled down", "3" representing a "neutral setting", and "5" was "mouth corners pulled up". As a visual shortcut, emoticons were used: "1" with the symbol ☹, "3" with 😐, and "5" with ☺. We cannot exclude associations with emotions, although we avoided giving direct hints to emotional states by mentioning terms such as sadness or happiness. We do not address the question whether the subjects perceived felt vs. non-felt smiles (cf. [4, 2]).

4 Results

Figure 3 provides an overview of the mean values for the eight possible versions of each vowel.

Comparing the "neutral" baselines (NNN) of the four vowels, /iː/ shows to be the most smiley-like and /yː/ the least smiley-like. In general, the stimuli with the highest scores are those with a higher F_0 (_ _ H). The best score was reached with a combination of all three parameters (SRH). This was true for all vowels except /uː/ where the combination with neutral lips scored best.

ANOVAs reveal significant differences for "spread lips" for /aː/ and /yː/ but not for /iː/, as expected because of the /iː/'s inherent setting of spread lips. /uː/ shows significantly less smileyness with spread lips.

"Raised larynx" causes a significant effect for /aː/ and /iː/ but not for /yː/ and /uː/. Here, for both rounded vowels a shortening of the vocal tract does not lead to a more smiley perception.

A different picture arises with "higher F_0": It causes significant differences on the perceptual smiley-scale for all four vowels.

5 Discussion and Future Work

Regarding the articulatory speech synthesizer used, we found that clear improvements can be achieved for all vowels investigated here. However, results suggest that it is not sufficient to exclusively manipulate lip spreading, larynx height (and voice quality), and F_0 in the way we presented here.

The overall effect of the F_0 parameter might indicate that our choice of neutral vs. raised F_0 was a too obvious manipulation. We are trained in everyday communication to detect even small F_0 changes to gather the intonation contour from a speech signal. Maybe further experiments should use smaller F_0 manipulations, i.e. more intermediate values, to match the subtleness of the lip and larynx manipulations. A control experiment could make sure that the participants did not "learn" to assess F_0 instead of smileyness.

Unrounded vowels can reach a higher smiley-score than rounded vowels when changing all three parameters. For the (in German) extremely rounded vowel /uː/, lip spreading without raising the fundamental frequency has to be examined further to see whether the low scores are found for *words* as well. /uː/ stimuli with spread lips always received lower scores than their counterparts with neutral lips, i.e. when larynx and F_0 parameters where kept constant. Obviously, an "injury" of the roundedness weighs more than a possible signal of smileyness. Or, perhaps, this vowel quality (close, back, *un*rounded) indicates an association with disgust for our listeners. The hypothesis here is that humans do not use (regular) lip spreading to achieve perceived smileyness but something else. It possibly is a combination of lip spreading and a reduction of mouth aperture at the same time by pressing together the far ends of the lips on each side – something which cannot be imitated with the current model of the synthesizer.

A general difficulty in the interpretation of the results is based on the fact that vowel quality is not very well preserved for some manipulated versions of /uː/ and /yː/. This might have led to confusion in the participants as to which abstract vowel they were listening to, to then be able to judge whether that assumed vowel was "smiled" or not. Another restriction is that the perception of stationary vowels might not be comparable to the perception of fluent smiled speech. Smileyness in fluent speech is probably not always at its extremes – possibly depending on the changing levels of emotional states as well as different sound categories: Some phones are possibly more exploited to convey smileyness than others. Dynamic changes *within* a vowel have already been shown to facilitate the perception of emotions [13].

Robson and MackenzieBeck [8] observed a more "i-face"-like articulation for open vowels, i.e. the vowels have a reduced jaw opening angle. Experiments with regard to reduced jaw opening for open vowels are necessary because they should be taken as a prerequisite to model *all* vowels for smiled ar-

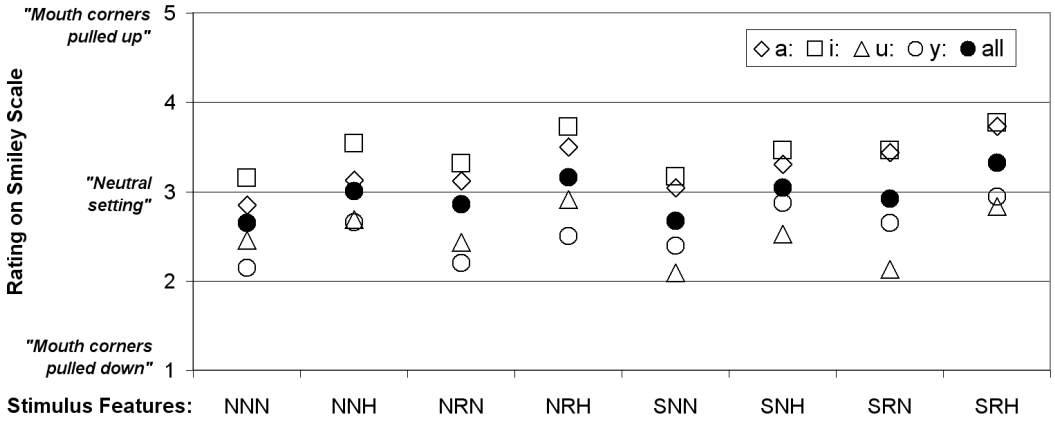

Figure 3: *Mean values of the smileyness of the four vowels. N = Neutral, S = Spread lips, R = Raised larynx, H = Higher* F_0.

ticulatory synthesized speech. However for consonants, especially those with labial activity such as [m, p, b, v, f, w], the changes remain unknown.

The main *visual* feature of human smiling (felt and non-felt) is lip spreading (cf. [4]). Supporting [10, 11, 8], our results raise the assumption that the auditory feature of smiling during articulation involves more than just a horizontal retraction of the lips and sometimes even avoids lip spreading.

Participant feedback showed that the smiley scale itself apparently invoked emotions for some of the subjects, interpreting the upper end of the scale (5) as "friendly" and commenting: You could also be "friendly" whilst speaking with "mouth corners pulled down". This mismatch or interference of dimensions has to be considered in future experiments.

Thus, extensions in future experiments could be to use longer utterances as test stimuli than just stationary vowels, to also integrate parameters like jaw opening angle, and to also use dynamic changes of parameters as applied in [13]. It is also advisable to integrate perceived phone quality into the perception test directly in order to use it as an additional variable for analysis.

References

[1] P. Birkholz and B. J. Kröger. Vocal tract model adaptation using magnetic resonance imaging. In *Proc. 7th ISSP, Ubatuba*, pages 493–500, 2006.

[2] A. Drahota, A. Costall, and V. Reddy. The vocal communication of different kinds of smile. *Speech Communication*, 51(4):278–287, 2008.

[3] J. A. Edmondson, J. Esling, J. G. Harris, L. Shaoni, and L. Ziwo. The aryepiglottic folds and voice quality in the Yi and Bai languages: Laryngoscopic case studies. *Mon Khmer Studies*, 31:83–100, 1999.

[4] P. Ekman and W. V. Friesen. *The Facial Action Coding System (FACS): A technique for the measurement of facial action*. Palo Alto, CA: Consulting Psychologists Press, 1978.

[5] E. Lasarcyk. Investigating larynx height with an articulatory speech synthesizer. In *Proc. 16th ICPhS, Saarbrücken*, pages 2213–2216, 2007.

[6] J. J. Ohala. An ethological perspective on common cross-language utilization of F0 in voice. *Phonetica*, 41:1–16, 1983.

[7] J. S. Perkell. *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*. Cambridge, MA, 1969.

[8] J. Robson and J. MackenzieBeck. Hearing smiles – Perceptual, acoustic and production aspects of labial spreading. In *Proc. 14th ICPhS, San Francisco*, pages 219–222, 1999.

[9] M. Schröder, V. Auberge, and M.-A. Cathiard. Can we hear smile? In *Proc. ICSLP, Sydney*, 1998. paper 0439.

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

[11] V. C. Tartter and D. Brown. Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America*, 96(4):2101–2107, 1994.

[12] Webpage. Synthesized vowel stimuli. http://www.coli.uni-saarland.de/∼evaly/issp08/.

[13] Y. Xu and S. Chuenwattanapranithi. Perceiving anger and joy in speech through the size code. In *Proc. 16th ICPhS, Saarbrücken*, pages 2105–2108, 2007.