# Perception of smiling in speech in different modalities by native vs. non-native speakers

*Caroline Émond* [1], *Albert Rilliard* [1], *Jürgen Trouvain* [2]

[1] LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
[2] Phonetics, Saarland University, Saarbrücken, Germany

caroemond@hotmail.com, albert.rilliard@limsi.fr, trouvain@coli.uni-saarland.de

## Abstract

Smiling, as a visual expression and nonverbal behavior, has been the subject of many studies, but less is known about smiled speech. This paper aims at examining the perception of smiling in audio only, visual only and in audiovisual conditions, among three different linguistic groups. The subjects' reaction times and the perceived intensity of the smiles were recorded, during a task where subjects rated stimuli for being or not produced with a smile. In order to proceed with an instrumental analysis, a Québec-French-speaking actress reproduced 138 utterances from spontaneous-speech data which served as stimuli for a perception test administered to French listeners from Québec (n=20) and France (n=18) and German listeners without any knowledge of Québec French (n=21). Results show that Québec listeners perceived a higher rate of smiling utterances followed by French and German. The reaction times are longer in audio only condition than in audiovisual. Listeners showed shorter reaction times for utterances that were associated with high-intensity as opposed to low-intensity smiling.

**Index Terms**: smiling in spontaneous speech, speech modalities, cross-cultural perception, emotional prosody

## 1. Introduction

In the context of everyday social life, it is difficult to imagine a day without smiling. Though smiling, as a visual expression and nonverbal behavior has been the subject of many studies [1, 2, 3], less is known about smiling as an audible expression occurring during speech. A smile requested to be produced during speech is known to be audible from the studies by Tartter [4, 5]. Similarly, utterances produced with emotionally neutral "spread lips" are perceived as being more "smiled" by listeners than those without lip spreading [6]. Such perceived smiles in speech are compared to non-smiled speech marked by higher formants (mostly F2 and F3) according to [4, 5, 6] and a higher F0 due to a higher overall muscular tension [4, p. 27]. Since then, several studies have focused on the production and perception of different kinds of "smiles" [7, 8, 9] or on the distinction between laughing, smiling, and crying speech [10]. Moreover, the relative weight of the audio and visual modalities in perception, and their interactions during audiovisual perception, is not well understood. In a global perception task of audiovisual stimuli, visual cues interfere with the auditory modality, as demonstrated by [8]. Our recent work examined universal versus culture-specific prosodic cues related to smiled speech [11] and the role of the gender of speakers and listeners in the perception of smiled speech [12, 13, 14, 15].

Most of the previous studies used various forms of controlled "lab speech" (as opposed to spontaneous data from "real life") and reading tasks, except for the study by Erickson et al. [10]. Moreover, data were produced in a monological context, involving a speaker who was not interacting with a listener. Even if the use of real-life data leads to several disadvantages when it comes to performing an instrumental analysis, the interactive structure (turn-taking organization, presence of discourse markers, etc.) of real-life data and the absence of unusual tasks in the lab may lead to the discovery of phenomena that cannot be directly observed in highly controlled speech (cf. e.g. [16]). This is why we have chosen to use spontaneous audio data as a first step, and then, as a second step, to record the same sentences with the help of an actress in order to perform acoustical analysis (not described here), and to have the video recordings. The general questions underlying this approach are twofold: When do naturally occurring speech utterances sound "smiled" (and what kind of conceptual interpretations may one give to these perceived "smiles")? Which cues (acoustic and visual) are related to the perception of this smiled speech?

In this paper, we present the results of a perception experiment designed to evaluate: the weight of the audio and visual modalities in the perception of smiling in speech, and a potential effect of the cultural and linguistic background of the subjects on the perception of smiling. Smiled speech is here opposed to speech-laughs [17, 18, 19]. This distinction is important considering that smiling and laughter are sometimes considered to be the same expression. A speech-laugh is a reinforcement of the expiratory activity of speech (e.g. stronger aspiration during unvoiced speech or a tremor during voice speech segments), often produced over a span of two syllables. In contrast, smiled speech is not produced with a breathy voice quality but often with a higher pitch and the impression of spread lips. Smiled speech usually takes a time span longer than two syllables.

## 2. Method

### 2.1. Stimuli and recordings

In preliminary studies, real-life spontaneous data of one female native speaker of Québec French [20] was used in order to investigate the ability of listeners in identifying smiled-speech and in evaluating its prosodic correlates (see [13, 14, 15] for details). Though acceptable for perception, the

real-life recordings were too noisy for a solid acoustic analysis.

For the current study, a professional actress (40 years old) with Québec French as native language was recorded while reproducing all the utterances that were extracted from the original corpus (69 smiling and 69 non-smiling utterances). Because it is impossible with spontaneous data to have exact non-smiling counterparts of smiling utterances, 69 utterances in a "neutral" condition were determined, which had an equivalent number of syllables (1 to 17 syllables, mean of 5.65 syllables) and average length (0.49 to 5.15 seconds) compared with the smiling utterances

The subset of the corpus was selected by the first author, a trained listener, for the perceptual experiment. To avoid an experimenter bias, the utterances perceived degree of smile was externally evaluated (the perception experiment). The non-smiling utterances were necessary to present differences in the acoustic parameters between the two conditions for further analyses.

These utterances were used for the analysis of the perception and production of smiling in audio only (AO), visual only (VO) and audiovisual (AV) conditions. The actress was told not to imitate or caricature, but to reproduce the intention and attitude that the female speaker was conveying with her voice. To ensure a high fidelity to the original data, each of the original utterances was played before every utterance recording.

The recordings took place in the soundproof booth of the Laboratoire de phonétique of Université du Québec à Montréal. A unidirectional stand-alone microphone (Shure SM86) at a stable distance of 60 cm from the mouth of the actress and a close-up of the face with a camera (Panasonic AG-DVX100bp) were used.

### 2.2. Participants

Fifty-nine listeners with no language, speech or hearing problems from three different linguistic groups were recruited at different universities for a perceptual experiment: the native French speakers came either from Quebec (QC – 4 males, 16 females, mean age: 24.8), or from France (FR – 12 males, 6 females, mean age: 27). The native German speakers were from Germany (DE – 6 males, 15 females, mean age: 25.8). Very few German speakers had a good knowledge of French from France, but none have heard Québec French before. Following debriefing after the test, they all report that they did not understand anything.

### 2.3. Procedures

The Parsour software [21] was used for performing the listening test and participants were presented with utterances in the three conditions of modality in the following order: AO – VO – AV. In AO and AV the participants were wearing headphones.

The task was exactly the same for each condition. The participants were instructed to determine if the utterances they heard, seen, and heard-and-seen were *smiling* or *not smiling* by clicking, as quickly as possible, on the appropriate emoticon (😐 or ☺) on the screen of a laptop. The *reaction times* were measured between the end of the audio/video file up to the mouse click on the emoticon associated with the response. If the utterances were perceived as smiling, the participants had to indicate the *intensity of the perceived smiles* by adjusting a visual analog scale on the computer screen (which consisted of a line with a minus sign (–) on the left-hand side and a plus sign (+) on the right-hand sign). The participants were presented with utterances in a random order and could listen to each of them only once. They had to click on "Poursuivre" ("to continue") when they were ready to evaluate the following utterance. A familiarization task (consisting in 8 utterances that could be heard as often as possible) with explanations preceded each part of the experiment. The entire test took about 60 minutes.

We consider a stimulus utterance as "smiled speech" if it was perceived as such by a majority of 75% of participants across the three different linguistic groups.

### 2.4. Data analysis

Responses with outlier reaction times (+/– 2.5 standard deviations) and reaction times under 100 milliseconds (ms) were excluded from the analysis (9.5% of the answers). This threshold is in line with [22] (cited in [23, p. 476]) who demonstrated that "genuine reaction times have a minimum value of at least 100 [milliseconds, which corresponds to the] time needed for physiological processes such as stimulus perception and for motor responses". The *smiling* or *non-smiling* answers and the intensity judgments for these outlier reaction times were not included in the data analysis.

Intensity ratings recorded by the Parsour software from the visual analog scale ranging from 1 (on the far left) to 100 (on the far right). These values were standardized for each subject and modality in z-scores and then divided into 5 categories of smiling intensity: "slightly," "somewhat," "moderately," "quite," and "very intense" – according to the following criterion: The intensity categories were computed as follows: Let $x$ be, for a language group and a modality, the median of the absolute values of the z-scores; the 5 categories are respectively: below -3$x$, between -3$x$ and –$x$, between –$x$ and +$x$, between +$x$ and +3$x$, and above +3$x$. Non-smiling sentences where given the category of smile intensity "no".

Percentage of smiling was analyzed as the proportion of "smiling" answers received by each sentence, using a logistic regression (based on R's *glm*() function [24], cf. [25, p. 633]). The optimal model contains the following factors: the modality (with three levels: AO, VO, AV), the linguistic group (three levels: QC, FR, DE), and the presented sentence (138 levels), plus the three double-interactions between these factors.

An analysis of reaction times variance was based on a linear mixed-effects model (using R's lmer() function [26]). The optimal model is based on four fixed effects (the modality of presentation, the linguistic group, the sentence, and the intensity category as defined above), and a random effect (the subject), nested in the linguistic group. The interactions between modality*intensity, modality*linguistic, modality* sentence, intensity*linguistic and modality*intensity* linguistic are used in the model. All factors, but the single effect of the linguistic group, have a highly significant effect on the reaction time.

## 3. Results

### 3.1. Percentage of perceived smiles

The analysis revealed that there is no effect of modality on the perception of smiling: participants perceived a comparable

average rate in AO, VO and AV conditions. All the other factors are highly significant, including the interactions. Thus, there is a significant effect of modality on the perception of individual sentences. About the effect of language, Québécois perceived a higher rate of smiling utterances followed by French people, and Germans as it can be seen in Figure 1. Despite limited amplitude in these differences, the language effect is consistent and significant.

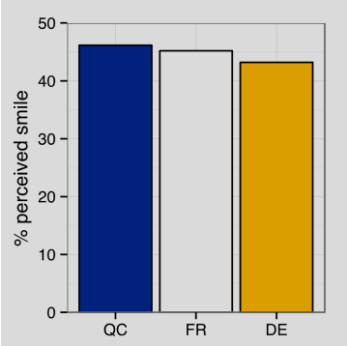


Figure 1: *Percentage of perceived smiles according to the linguistic group.*

The significant interaction between the linguistic group and the modality (Figure 2) shows that Germans tend to perceive more smiles in AO and fewer in VO and AV compared to Québécois, the perception rate of French being quite close to the one of Québécois.

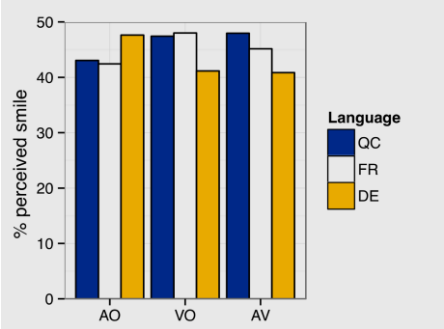


Figure 2: *Percentage of perceived smiles by modality according to the linguistic group.*

### 3.2. Reaction times

For reaction time, the random effect of *subjects* explains 27% of the variance. The effect of modality on reaction times (RT) shows that the AO condition leads to longer decision times, contrary to the AV condition as illustrated in Figure 3.

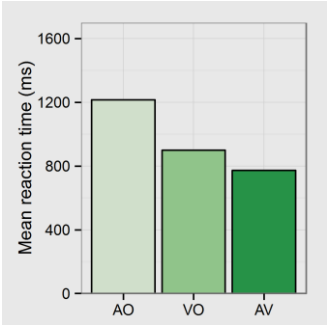


Figure 3: *Mean RTs for AO, VO and AV modalities.*

Figure 4 shows that participants had faster reaction times for utterances associated with high smile intensity, as opposed to low intensity. In other words, the more intense the smile was perceived, the quicker the answer. The same result was observed by [13] and [15] in AO.

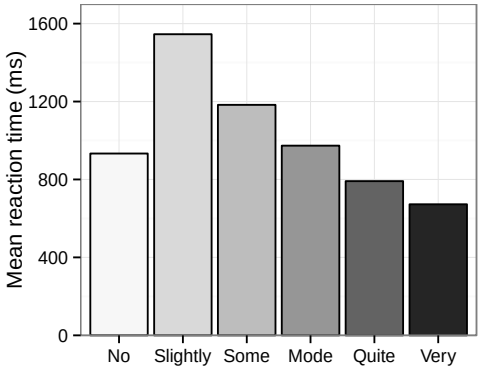


Figure 4: *Mean reaction times by perceived intensity (No = non-smiling utterances, Some = somewhat, Mode = moderately).*

The relation between reaction times and intensity categories holds for the different modalities and linguistic groups, as it can be seen in Figure 5. Specifically, Figure 5A shows that the "slightly intense smiling" category has slower reaction times for French and Germans in AO. This difference is however more salient for French. In VO, like in AO, the French group took significantly more time to make their decisions when a smile is perceived as "slightly smiling" (Figure 5B). Figure 5C shows that the difference of processing time for the "slightly smiling" category is not replicated in the AV modality for the French group: in the AV modality, there is no effect of the group on the reaction time.

## 4. Discussion

Not surprisingly, and as expected and demonstrated by many researchers, smiling is audible when produced synchronously with speech. Listeners from another native language variety also perceive a similar average rate of smiling expressions, whatever the modality (between 40 to 50%, while 50% of the presented stimuli were supposed to be performed with smile). We have seen that the perception rate of smiling given by Québécois is followed by those of French and then those of Germans. These small but significant differences indicate that the further apart listeners are from the source language, the lower is the percentage of perceived smiling. In [15] we showed that the semantic content of these utterances did not bias native speakers toward the perception of smiling. With this in mind, the present result suggests that this language group difference is related to prosodic and visual cues, i.e. culture-specific cues.

Whatever the modality, Québécois and French behave in much the same manner, perceiving more smiling utterances in VO and AV than in AO. For these two groups, the facial expression seems to carry more information about smiling than did the voice in the AO condition. The pattern for Germans is opposed, as they perceived more smiles in AO and fewer in VO and AV. For the Germans it might be that the linguistic 'blindness' helped them to concentrate on the voice quality in the AO condition but that the visual information did

not match. Thus, the information carried by the audio modality could be related to the language while the information in the visual modality is linked to other culture-specific attitudes – e.g. different kinds of smiles that are not related to amusement or enjoyment for example.

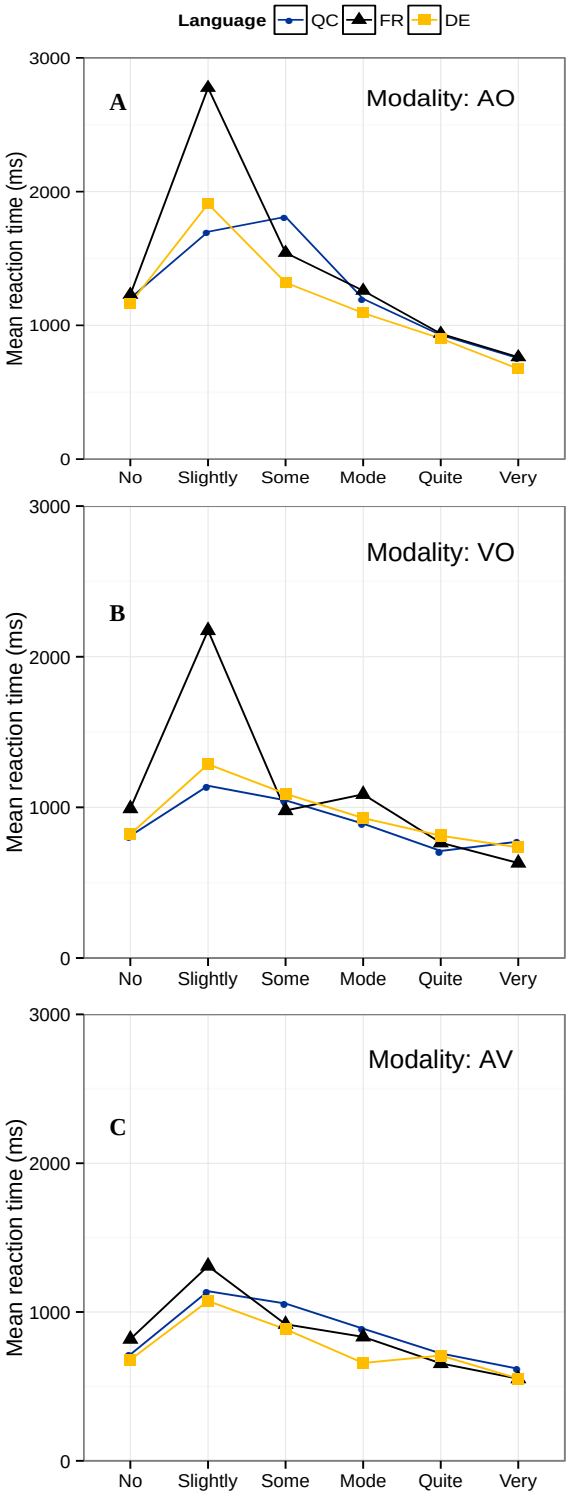


Figure 5: *Mean reaction times by perceived intensity for each linguistic group in (A) the AO, (B) the VO, (C) the AV conditions (No= non-smiling, Some= somewhat, Mode=moderately).*

Reaction times were also investigated and used as a tool to evaluate the speed of processing of smiling for different modalities and intensity categories of the perceived smile. The results showed that participants made faster decisions (had quicker reaction times) in AV followed by VO and then AO. Having both audio and visual information leads to quicker reaction times in the perception of smiling by the three language groups.

Regardless of the modality of presentation and the origin of the participants, utterances perceived with a "very intense smiling" had shorter reaction times than utterances perceived as "slightly intense". In other words, reaction times increase as the intensity of the perceived smile decreases. This suggests that prototypical smiles are characterized by a high-intensity level and are perceived more quickly. The pattern is nearly the same by modality for the three groups, apart for French who had longer reaction times than Québécois and Germans in the "slightly intense" smiling category in AO and VO. Here, the possibility of having an influence of the language variety cannot be excluded when only one modality is presented. A more subtle expression can contain or combine other attitudes or emotions. The semantic content does not necessarily influence their decision but maybe French tried to find cues in it in order to decide on the "smileyness" of low-intensity utterances. This effect disappears in AV modality.

The results obtained in [15] with spontaneous real-life data are validated here with the same corpus but interpreted by an actress. The principal limit is about the comparison that cannot be make between identical smiling and non-smiling utterances since it is impossible to control real-life data.

## 5. Conclusion

On the whole, the source language does not seem to have a considerable effect on the perception of smiling, at least where the linguistic groups share or are exposed to several similar cultural influences (e.g. western countries as opposed to eastern countries). The audiovisual modality provides more information which help to determine the smiling nature of a given utterance. Prototypical smiles tend to be perceived with a high-intensity level as they are relatively quickly identified by participants.

The next step is the analysis of acoustical data, which is currently in progress. We should also address the question of the various types of smiles in conversation and their acoustic characteristics. This could ultimately help developing new methods for a better understanding of the important affective and social signal of smiling from an acoustic and prosodic perspective.

## 6. Acknowledgements

This work was supported by a postdoctoral grant from the Social Sciences and Humanities Research Council (SSHRC). A great big thank to Lucie Ménard, Paméla Trudeau-Fisette and other members of the Laboratoire de phonétique at the Université du Québec à Montréal. This work would not have been possible without the participation of many persons from Québec, France and Germany.

# 7. References


[1] P. Ekman, R. J. Davidson and W. V. Friesen, "The Duchenne Smile: Emotional Expression and Brain Physiology II", *Journal of Personality and Social Psychology*, vol. 58, no. 2, pp. 342-353, 1990.

[2] M. H. Abel, Ed., *An Empirical Reflection of Smile*, Lewiston: Edwin Mellen Press, 2002.

[3] M. Méhu, "Smiling and Laughter in Naturally Occurring Dyadic Interactions: Relationship to Conversation, Body Contacts, and Displacement Activities", *Human Ethology Bulletin*, vol. 26, no. 1, pp. 10-28, 2011.

[4] V. C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech", *Perception and Psychophysics*, vol. 27, no. 1, pp. 24-27, 1980.

[5] V. C. Tartter and D. Braun, "Hearing smiles and frowns in normal and whisper registers", *Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2101-2107, 1994.

[6] J. Robson, and J. Mackenzie Beck, "Hearing smiles – Perceptual, acoustic and production aspects of labial spreading", *Proceedings of 14th ICPhS*, pp. 219-222, 1999.

[7] M. Schröder, V. Aubergé and M.-A. Cathiard, "Can we hear smile?", *Proceedings of the Conference on Spoken Language Processing*, vol. 3, pp. 559-562, 1998.

[8] V. Aubergé and M. Cathiard, "Can we hear the prosody of smile?", *Speech Communication*, vol. 40, no. 1-2, pp. 87-97, 2003.

[9] A. Drahota, A. Costall and V. Reddy, "The vocal communication of different kinds of smile", *Speech Communication*, vol. 50, no. 4, pp. 278-287, 2008.

[10] D. Erickson, C. Menezes and K. Sakakibara, K., "Are you laughing, smiling or crying?", *Proceedings of 2009 APSIPA Summit and Conference*, pp. 529-537, 2009.

[11] C. Émond, J. Trouvain and L. Ménard, "Perception of French smiled speech by native vs. non-native listeners: a pilot study", *Proceedings of the Interdisciplinarity Workshop on the Phonetics of Laughter – 16th ICPhS*, pp. 27-30, 2007.

[12] C. Émond, *Les corrélats prosodiques et segmentaux de la parole souriante en français québécois*, Master Thesis, Université du Québec à Montréal, 2008.

[13] C. Émond and M. Laforest, "Prosodic correlates of smiled speech", *Proceedings of Meetings on Acoustics*, vol. 19, 060220, 2013.

[14] C. Émond, L. Ménard and M. Laforest, "Perceived correlates of smiled speech in spontaneous data", *Proceedings of Interspeech*, 2013.

[15] C. Émond, *Les corrélats prosodiques et fonctionnels de la parole perçue souriante en français québécois spontané*, PhD Thesis, Université du Québec à Montréal, 2014.

[16] P. Wagner, J. Trouvain and F. Zimmerer, "In defense of stylistic diversity in speech research", *Journal of Phonetics*, vol. 48, pp. 1-12, 2015.

[17] E. E. Nwokah, H.-C. Hsu, P. Davies and A. Fogel, "The Integration of Laughter and Speech in Vocal Communication: A dynamic Systems Perspective", *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 4, pp. 880-894, 1999.

[18] J. Trouvain, "Phonetic Aspects of 'Speech-Laughs'", *Proceedings of ORAGE, Orality and Gestuality Conference*, pp. 634-639, 2001.

[19] J. Trouvain, "Segmenting Phonetic Units in Laughter", *Proceedings of 15th ICPhS*, pp. 2793-2796, 2003.

[20] D. Vincent, M. Laforest and G. Martel, "Le corpus de Montréal 1995. Adaptation de la méthodologie sociolinguistique pour l'analyse conversationnelle", *Dialangue*, vol. 6, pp. 29-45, 1995.

[21] M. Bastien, C, Émond and L. Ménard, L., *Parsour 1.60*, http://microbe.ca/, 2015.

[22] R. D. Luce, *Response times: Their role in inferring elementary mental organization*, New York: Oxford University Press, 1986.

[23] R. Whelan, "Effective analysis of reaction time data", *The Psychological Record*, vol. 58, no. 3, pp. 475-482, 2008.

[24] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[25] M. J. Crawley, *The R book*, Chichester: Wiley, 2013.

[26] D. Bates, M. Mächler, B. Bolker and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1-48, 2015.