Uncovering mental representations of smiled speech using reverse correlation

Emmanuel Ponsot, Pablo Arias and Jean-Julien Aucouturier

Citation: The Journal of the Acoustical Society of America **143**, EL19 (2018); doi: 10.1121/1.5020989
View online: https://doi.org/10.1121/1.5020989
View Table of Contents: https://asa.scitation.org/toc/jas/143/1
Published by the Acoustical Society of America

Uncovering mental representations of smiled speech using reverse correlation

Emmanuel Ponsot,[a] Pablo Arias, and Jean-Julien Aucouturier
STMS (Sciences et Technologies de la Musique et du Son) Lab (Ircam/CNRS/UPMC),
1 place Igor Stravinsky, Paris, France
ponsot@ircam.fr, arias@ircam.fr, aucouturier@ircam.fr

Abstract: Which spectral cues underlie the perceptual processing of smiles in speech? Here, the question was addressed using reverse-correlation in the case of the isolated vowel [a]. Listeners were presented with hundreds of pairs of utterances with randomly manipulated spectral characteristics and were asked to indicate, in each pair, which was the most smiling. The analyses revealed that they relied on robust spectral representations that specifically encoded vowel's formants. These findings demonstrate the causal role played by formants in the perception of smile. Overall, this paper suggests a general method to estimate the spectral bases of high-level (e.g., emotional/social/paralinguistic) speech representations.

[DDO]
Date Received: July 28, 2017 **Date Accepted:** December 24, 2017

1. Introduction

Human social interaction relies, for the most part, on the ability to extract and decode facial and vocal expressions, from which we can infer each other's social traits and emotional states (Willis and Todorov, 2006). Among such expressions, the smile—the bilateral stretching of the lips by the zygomaticus major muscles—is remarkable for being produced and recognised early in development (Oostenbroek *et al.*, 2016) and across cultures (Ekman *et al.*, 1969). Smiling is not only perceived visually, but can also be heard in spoken voice (or smiled speech, Tartter, 1980; Basso and Oullier, 2010). Although the acoustic consequences of smiling during speech production have been studied from recorded speech corpora (Barthel and Quéné, 2015; Podesva *et al.*, 2015), little is known about how the human auditory system processes such cues. Here, we used a psychophysical reverse-correlation approach to explore what makes a voice sound smiled on the basis of its spectral cues.

The general idea of reverse-correlation is to present a system (here, the human observer) with a slightly perturbed stimulus over many trials. This perturbation is created by directly adding white noise to a stimulus or by manipulating higher-level dimensions using random deviations around baseline. Perturbed stimuli will, on different trials, lead to different responses of the system, and the tools of reverse-correlation can be used to infer the functional properties of the system from the pattern of stimulus noise and their associated responses. The technique was first used by psychophysicists to characterize human sensory processing (e.g., detection of tones in noise; Ahumada and Lovel, 1971; discrimination of frequency distributions; Berg, 1989) but it is also a powerful tool to characterize higher-level perceptual or cognitive processes, for which it can uncover the "optimal stimulus" (or "mental representation") that is driving participant responses.

In vision, reverse-correlation was applied to derive observers' mental representations of, e.g., what makes a face happy (Mangini and Biederman, 2004), how facial expressions differ across cultures (Jack *et al.*, 2012a) or even what makes Mona Lisa seem to smile (Kontsevich and Tyler, 2004). A few recent studies have started to use the approach for auditory tasks such as speech intelligibility (Varnet *et al.*, 2016; Venezia *et al.*, 2016) or musical instrument recognition (Thoret *et al.*, 2016). In particular, Brimijoin *et al.* (2013) have used reverse-correlation to uncover the internal representations of a whispered vowel by presenting random-spectrum static noises to human listeners. Their results showed that humans possess strikingly fine spectral mental

[a] Author to whom correspondence should be addressed. Also at: Laboratoire des Systèmes Perceptifs (CNRS UMR 8248) and Département d'études cognitives, Ecole Normale Supérieure, PSL Research University, Paris, France.

representations of a vowel, with spectral weights aligned to the formant frequencies of real whispered vowels.

In the present study, we used reverse correlation to characterize the perceptual filters employed by humans to infer whether a person is smiling from the spectral characteristics of the voice, in particular in the vowel [a] pronounced by a male speaker. We also assessed the robustness of listeners' perceptual decoding in the task by quantifying their internal noise using a double-pass procedure.

2. Materials and methods

2.1 Ethics

The protocol of this experiment was approved with an IRB given by the "Institut Européen d'Administration des Affaires" (INSEAD).

2.2 Subjects

Ten participants (5 women, 5 men; age 18–29 yrs) were recruited for the experiment. None reported having hearing problems. In accordance with APA Ethical Guidelines, participants gave their informed written consent prior to the experiment and were debriefed about the true purpose of the research immediately after. Participants were paid for their participation.

2.3 Stimuli

We recorded an utterance of the phoneme [a], pronounced with constant pitch (~122 Hz) by a single male speaker with a neutral facial expression (Mm. 1), and selected a 500-ms stationary part of the sound to create a stimulus with constant spectral energy. We then produced many spectral variants of this baseline stimulus by manipulating its spectral characteristics using a random frequency equalizer composed of 25 linearly interpolated, log-separated frequency points spaced between 100 and 10 000 Hz, with gain values (in dB) drawn from Gaussian distributions [standard deviation (SD) = 5 dB clipped at ±2.5 SD].

Mm. 1. Audio file of the original /a/ vowel (pronounced with a neutral facial expression). This is a file of type "wav" (44 Ko).

2.4 Apparatus

All stimuli were mono sound files generated at a sampling rate of 44.1 kHz with 16-bit resolution using MATLAB. They were presented diotically through headphones (Beyerdynamic DT 770 PRO, 80 ohms) at the same level for all participants (~70 dB sound pressure level). Sound levels were measured using a Brüel & Kjær 2238 Mediator sound-level meter placed at a distance of 4 cm from the right (left) earphone. A DPA 4066 omni-directional microphone was used to record the voice of the male speaker employed to create the stimuli.

2.5 Procedure

The experiment consisted of a single 1 h experimental session including 6 blocks of 100 trials. Using a 2I, 2AFC procedure, participants were presented pairs of randomly-filtered voices (example: Mm. 2) and asked in each pair which of the two appeared to have been produced with the greatest smile. Since there were no correct or incorrect answers, participants did not receive trial-by-trial feedback. Trials presented in the first 5 blocks were all different, but the 100 trials of the sixth block were the same as those presented in the fifth block (in the same order). This double-pass procedure was used to evaluate the percentage of agreement and thus the level of internal noise for each observer in the task. None of the observers noticed this repetition.

Mm. 2. Audio file of a trial presented in the experiment. This is a file of type "wav" (132 Ko).

2.6 Data analysis

One reverse-correlated frequency filter (a 25-points vector) was computed for each subject as the mean filter of the voices classified as smiling from which we subtracted the mean filter of the remaining voices that were not chosen as smiling by the participant (the data collected during the sixth block, i.e., the same trials as in the fifth block, were not used to derive these filters).

Formant frequencies and bandwidths were computed using Praat (Boersma and Petrus, 2002). The spectral envelopes were extracted using the true envelope implementation of IRCAM's Super-VP tool (Villavicencio *et al.*, 2006). Formant gains were estimated as the values of the spectral envelope at the formant frequencies.

3. Results

3.1 Perceptual filters and mental prototypes

The averaged reverse-correlated frequency filter underlying the evaluation of smile in the [a] vowel used in the task is plotted in Fig. 1(a). This filter presents clear structures aligned with the formant frequencies and bandwidths of the original phoneme and an overall enhancement of the high frequencies compared to the low frequencies. Because the reverse-correlation technique only allows the derivation of participants' internal filters with amplitudes that are proportional to the SD of the external noise used in the experiment, we derived prototype stimuli for smiling and non-smiling voices by applying the filters to the base stimulus with a gain of ±50 [Fig. 1(b)], a value chosen to qualitatively match averaged spectral-envelope differences of the stimuli presented in the experiment. These prototypes have fair intra-individual consistency and appear to implement distinctive operations on the formants: (i) formants 1 and 2 are represented with increased frequency in the smiling prototype (in red, Mm. 3), compared to the non-smiling prototype (in blue, Mm. 4) and (ii) formants 2, 3, and 4 are represented with increased amplitude. Figure 1(c) presents the difference between the spectral envelopes computed over these prototypes: it is virtually identical to the raw filter plotted in Fig. 1(a), showing that the filter profiles represent the real physical changes that occurred on spectral envelopes. Overall, as summarized in Fig. 1(d), the spectral transformations needed to perceive the phoneme as smiling consist primarily of a frequency increase of $F1$ and $F2$ and an amplification of $F2$, $F3$, and $F4$.

Mm. 3. Smiling audio prototype derived from the perceptual results. This is a file of type "wav" (44 Ko).

Mm. 4. Non-smiling audio prototype derived from the perceptual results. This is a file of type "wav" (44 Ko).

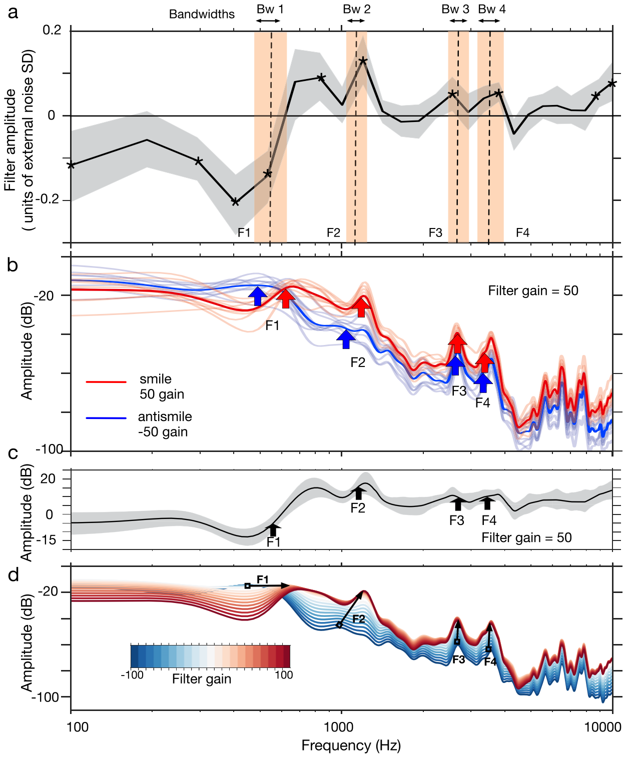

Fig. 1. (Color online) (a) Averaged filter underlying the judgment of the vocal smile, as derived with reverse-correlation. Asterisks indicate significant differences from 0 (two-tailed; paired-sample t-tests, $p < 0.05$). Vertical shaded areas indicate how the first four formants of the voice align with the structure of the filter. (b) When this filter or its opposite is applied (here with a gain of 50) to the original voice, it reveals the internal auditory representations of a smiling (Mm. 3) and a non-smiling voice (Mm. 4). Shaded lines represent the corresponding spectral envelope for each participant's internal filter. (c) Mean spectral envelope difference between smiling and non-smiling sounds for a filter gain of 50. (d) Mean spectral envelope across participants for different filter gains, highlighting the overall transformation over the spectral envelopes as one goes from mental representations of a strongly non-smiling voice, to that of a strongly smiling voice. Shaded areas represent 95% confidence intervals computed with a bootstrap procedure.

3.2 Observers' consistency

The double-pass methodology was used to assess observers' consistency from a measure of internal noise relative to the external noise added to the stimuli (i.e., the random spectral changes): the last two blocks were identical so that all observers received the same 100 trials twice. All but two participants (who had percentages of agreement of 49% and 51%) performed well above chance level over these repeated blocks: when these two participants were removed, the average percentage of the same responses over these two blocks was 68.1% (SD = 5.8). We then estimated the amount of internal noise for each of these remaining eight subjects using a simple signal detection theory model with late additive noise[1] (Burgess and Colborne, 1988; Neri, 2010). We found an average internal noise level of 1.2 (SD = 0.9), as expressed in units of external noise SD. There are no measures of internal noise in facial emotion or visual smiles reverse-correlation studies we can compare our estimate with, but it is of note that previously reported values in other high-level visual processing tasks are generally higher than our estimate; a value of 2 is found for human biological motion discrimination (running vs walking) (van Boxtel and Lu, 2015), an average value greater than 2 in the evaluation of ensemble average size (Im and Halberda, 2013) and values between 1 and 4 for face identification (Gold *et al.*, 2004). Our value rather corresponds to what has been found on average in many different low-level auditory and visual tasks (Neri, 2010), suggesting that the high-level auditory filtering of smile in speech relies on a fairly stable processing, i.e., that observers possess robust and stable auditory representations of what makes a smiling voice.

4. Discussion and conclusion

This paper examines the spectral filtering that underlies the auditory processing of smile in the human voice using behavioral reverse-correlation. We show that humans rely on robust mental representations that allow them to tell whether a voice is smiling or not, and that these internal representations can be assessed with high precision with our method. We computed the filters in the [100–10 000 Hz] range, and found that the spectral characteristics of a smiled [a] phoneme are mentally represented with increased $F1$ and $F2$ frequency, increased $F2$, $F3$ and $F4$ amplitude, and an overall enhancement of the high frequencies compared to the low frequencies. The structure of these filters demonstrates a delicate ability of the auditory system to parse amplitude and frequency changes by formant. Surprisingly, even though these acoustic transformations are complex and non-linear across the spectrum, internal representations were fairly consistent across participants, demonstrating that auditory consequences of articulatory gestures associated to smiling are accurately available even to naive participants. Indeed, listeners were able to robustly infer smile from subtle random changes in the spectral cues of a token.[2] The fact that the average level of internal noise was comparable to the one measured in the low-level auditory and visual tasks and that the underlying perceptual filters were finely tuned with formant modifications further suggests that observers rely on deep, robust auditory representations of what is smiled, and what is not.

Our results complement previous studies on vocal and synthetic productions. Recent studies have shown that smiling during speech production has several acoustic consequences. Increases in loudness, $F0$ and $F2$ (Barthel and Quéné, 2015; Podesva *et al.*, 2015), as well as higher $F1$ and $F2$ dispersions (Drahota *et al.*, 2008) were reported. Interestingly, although $F0$ changes are often found when analysing smiling vocal productions, these alterations do not seem to be necessary to recognize smiles in speech, as smiles can also be recognised in nonsense syllables without training (Tartter, 1980) and in whispered voices (Tartter and Braun, 1994). For the particular [a] phoneme, Keough *et al.* (2015) have reported an increase of $F1$ and $F2$ during production, which is in line with our findings. Other acoustical analyses of smiling vs neutral productions of speech showed that, even if the acoustic consequence of a smiling gesture on formants depends on the vowel (Barthel and Quené, 2015; Fagel, 2010; El Haddad *et al.*, 2015; Keough *et al.*, 2015), these always exhibit an overall increase in frequency. Thus, it can reasonably be assumed that the filters returned with our method for other vowels and/or speakers[3] would commonly implement changes on the formant structure of the tokens, but these would be specific to the phoneme considered. If such is the case, listeners' ability to recognise smiled phonemes would be a remarkable mechanism, as the acoustic consequences of smiling are non-linear across the spectrum and across phonemes. Future studies should provide a complete picture of the perceptual decoding of auditory smiles, e.g., by testing different phonemes from different speakers as well as transitions with consonants (CVs, VCs, or CVCs, e.g., using the paradigm employed in Varnet *et al.*, 2016).

All in all, the present results shed light onto remarkable abilities of the human auditory system to use the voice's acoustic features to infer an associated facial articulatory gesture, and are consistent with recent neuroimaging work showing that phonetic representations are encoded in the human auditory cortex (Mesgarani *et al.*, 2014). Whether the smile prototypes uncovered in this study are encoded in the auditory system or stem from information incoming from motor areas (Hickok *et al.*, 2011; Pulvermüller *et al.*, 2006) is a question that remains to be elucidated.

Smiling is a highly important behaviour in the emotional expressive repertoire. The present paradigm could be used to study other acoustical facets of smiled speech [e.g., are different types of smiles related to different representations, as is the case in vision? (Rychlowska *et al.*, 2017)], but also to explore the bases of other social or articulatory traits, and investigate whether other vocal gestures of lesser emotional or adaptive relevance are processed through similarly robust and consistent perceptual filters. The present paper suggests a general framework[4] to estimate the spectral bases of any high-level representation of speech, i.e., not only smiled speech, but any emotional/social/paralinguistic aspect of speech timbre.

Acknowledgments

We thank Peter Neri for his comments related to internal noise and Louise Vasa for her help conducting the experiment. The study was supported by ERC Starting Grant CREAM (Grant No. 335536, to J.-J.A.). Data collected at the Centre Multidisciplinaire des Sciences Comportementales Sorbonne Universités-INSEAD. E.P. and P.A. contributed equally to this work.

References and links

[1] As the two stimuli presented in each trial were not distinguishable from the point of view of signal detection theory, we used a model with common distribution to construct "signals" (smiling voices) and "noises" (non-smiling voices), i.e., corresponding to $d' = 0$ or percent correct of 0.5.

[2] Although such stimuli (with random equalisation) might sound synthetic, listeners were able to interpret them as speech and could employ their mental representations of smile to do the task.

[3] The choice of the [a] vowel was motivated by the fact that it is located in the middle of the phonological space (mid-opening of the mouth), therefore being highly compatible with the smiling gesture. This allowed us to efficiently deploy reverse-correlation without being constrained by the space boundaries, because spectral manipulations could make the token more or less smiling. It would have been difficult to measure perceptual filters for smiling with phonemes close to the boundaries of the space, i.e., that cannot easily be produced with a smiling/unsmiling face (e.g., an [i], produced with the jaw in close to smile position or a [u], produced with rounded lips antithetical to smiling). Similar considerations have been made in visual reverse-correlation studies requiring morphed unexpressive/androgyny faces (e.g., Jack *et al.*, 2012b).

[4] Cleese, the voice-manipulation software used in the study is open-source and made available at http://cream.ircam.fr/?p=521.

Ahumada, A. J., and Lovell, J. (**1971**). "Stimulus features in signal detection," J. Acoust. Soc. Am. **49**(6B), 1751–1756.

Barthel, H., and Quené, H. (**2015**). "Acoustic-phonetic properties of smiling revised–measurements on a natural video corpus," in *Proceedings of the 18th International Congress of Phonetic Sciences*, The University of Glasgow, Glasgow, United Kingdom.

Basso, F., and Oullier, O. (**2010**). " 'Smile down the phone': Extending the effects of smiles to vocal social interactions," Behav. Brain Sci. **33**(06), 435–436.

Berg, B. G. (**1989**). "Analysis of weights in multiple observation tasks," J. Acoust. Soc. Am. **86**(5), 1743–1746.

Boersma, P., and Petrus, G. (**2002**). "Praat, a system for doing phonetics by computer," Glot Int. **5**(9), 341–345.

Brimijoin, O. W., Akeroyd, M. A., Tilbury, E., and Porr, B. (**2013**). "The internal representation of vowel spectra investigated using behavioral response-triggered averaging," J. Acoust. Soc. Am. **133**(2), EL118–EL122.

Burgess, A. E., and Colborne, B. (**1988**). "Visual signal detection. IV. Observer inconsistency," J. Opt. Soc. Am. A **5**(4), 617–627.

Drahota, A., Costall, A., and Reddy, V. (**2008**). "The vocal communication of different kinds of smile," Speech Commun. **50**(4), 278–287.

Ekman, P., Sorenson, E. R., and Friesen, W. V. (**1969**). "Pan-cultural elements in facial displays of emotion," Sci. **164**(3875), 86–88.

El Haddad, K., Dupont, S., d'Alessandro, N., and Dutoit, T. (**2015**). "An HMM-based speech-smile synthesis system: An approach for amusement synthesis," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 5, pp. 1–6.

Fagel, S. (**2010**). "Effects of smiling on articulation: Lips, larynx and acoustics," in *Development of Multimodal Interfaces: Active Listening and Synchrony* (Springer, Heidelberg Dordrecht, London), pp. 294–303.

Gold, J. M., Sekuler, A. B., and Bennett, P. J. (**2004**). "Characterizing perceptual learning with external noise," Cognit. Sci. **28**(2), 167–207.

Hickok, G., Houde, J., and Rong, F. (**2011**). "Sensorimotor integration in speech processing: Computational basis and neural organization," Neuron **69**(3), 407–422.

Im, H. Y., and Halberda, J. (**2013**). "The effects of sampling and internal noise on the representation of ensemble average size," Attn., Percept., Psychophys. **75**(2), 278–286.

Jack, R. E., Caldara, R., and Schyns, P. G. (**2012b**). "Internal representations reveal cultural diversity in expectations of facial expressions of emotion," J. Exp. Psychol. **141**(1), 19–25.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., and Schyns, P. G. (**2012a**). "Facial expressions of emotion are not culturally universal," Proc. Natl. Acad. Sci. U.S.A. **109**(19), 7241–7244.

Keough, M., Ozburn, A., McClay, E. K., Schwan, M. D., Schellenberg, M., Akinbo, S., and Gick, B. (**2015**). "Acoustic and articulatory qualities of smiled speech," Can. Acoust. **43**(3), available at https://jcaa.caa-aca.ca/index.php/jcaa/article/view/2797.

Kontsevich, L. L., and Tyler, C. W. (**2004**). "What makes Mona Lisa smile?," Vis. Res. **44**(13), 1493–1498.

Mangini, M. C., and Biederman, I. (**2004**). "Making the ineffable explicit: Estimating the information employed for face classifications," Cognit. Sci. **28**(2), 209–226.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (**2014**). "Phonetic feature encoding in human superior temporal gyrus," Science **343**(6174), 1006–1010.

Neri, P. (**2010**). "How inherently noisy is human sensory processing?," Psychonomic Bull. Rev. **17**(6), 802–808.

Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Costantini, S., Davis, J., and Slaughter, V. (**2016**). "Comprehensive longitudinal study challenges the existence of neonatal imitation in humans," Curr. Biol. **26**(10), 1334–1338.

Podesva, R. J., Callier, P., Voigt, R., and Jurafsky, D. (**2015**). "The connection between smiling and GOAT fronting: Embodied affect in sociophonetic variation," in *Proceedings of the International Congress of Phonetic Sciences*, Vol. 18.

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., and Shtyrov, Y. (**2006**). "Motor cortex maps articulatory features of speech sounds," Proc. Natl. Acad. Sci. U.S.A. **103**(20), 7865–7870.

Rychlowska, M., Jack, R. E., Garrod, O. G., Schyns, P. G., Martin, J. D., and Niedenthal, P. M. (**2017**). "Functional smiles: Tools for love, sympathy, and war," Psychol. Sci. 1259–1270.

Tartter, V. C. (**1980**). "Happy talk: Perceptual and acoustic effects of smiling on speech," Attn., Percept., Psychophys. **27**(1), 24–27.

Tartter, V. C., and Braun, D. (**1994**). "Hearing smiles and frowns in normal and whisper registers," J. Acoust. Soc. Am. **96**(4), 2101–2107.

Thoret, E., Depalle, P., and McAdams, S. (**2016**). "Perceptually salient spectrotemporal modulations for recognition of sustained musical instruments," J. Acoust. Soc. Am. **140**(6), EL478–EL483.

van Boxtel, J. J. A., and Lu, H. (**2015**). "Joints and their relations as critical features in action discrimination: Evidence from a classification image method," J. Vis. **15**(1), 20.

Varnet, L., Meunier, F., Trollé, G., and Hoen, M. (**2016**). "Direct viewing of dyslexics' compensatory strategies in speech in noise using auditory classification images," PLoS One **11**(4), e0153781.

Venezia, J. H., Hickok, G., and Richards, V. M. (**2016**). "Auditory 'bubbles': Efficient classification of the spectrotemporal modulations essential for speech intelligibility," J. Acoust. Soc. Am. **140**(2), 1072–1088.

Villavicencio, F., Robel, A., and Rodet, X. (**2006**). "Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. I–I.

Willis, J., and Todorov, A. (**2006**). "First impressions making up your mind after a 100-ms exposure to a face," Psychol. Sci. **17**(7), 592–598.