To appear in Linguistics Vanguard

Smiles can be conveyed to listeners via speech signals: A case study with Japanese labiodentalization[1]

To appear in *Linguistics Vanguard*

Shigeto Kawahara (Keio University)

Kimi Akita (Nagoya University)

Zen Nagatsuka (Zen Voice Factory)

Julián Villegas (University of Aizu)

Abstract

One long-standing question in speech communication research concerns the extent to which an emotional state is encoded in the acoustic signal and conveyed to the listeners. To address this question, we deployed a recent finding in Japanese, in which speakers replace bilabial sounds (e.g. [b]) with labiodental sounds (e.g. [v]) when speaking with a smile. Experiment 1 used stimuli recorded by two professional voice actors under four conditions crossing two factors, Voice Quality (Smile vs. Non-Smile) and Articulation (Labiodental vs. Bilabial). The results confirm that labiodentals function as a reliable cue for smiling, even in the absence of visual information. Experiment 2 controlled for the effects of F0 and amplitude of the stimuli, and we still observed an effect of labiodentals. We conclude (1) that smiles can be conveyed via speech signals to the listeners without any visual cues and (2) that in the case of Japanese, this can be achieved via the use of labiodentals.

1 Introduction

Unlike written texts, speech conveys far more than linguistic information. As already noted in

[1] We thank the two voice actors (Ayaka Asai and Runa Morikawa) who offered their professional skills to produce the experimental stimuli and Mone Kamishiraishi for bringing this pattern of labiodentalization to our attention. This research is supported by the JSPS grant 25K04035. We received useful comments from two anonymous reviewers. Claude and Gemini were used to assist the write-up of this paper. The authors, however, reviewed and edited the final content and take full responsibility for the final product.

the foundational work on speech production and perception (e.g. Abercrombie 1967; Fujisaki 2004; Laver 1994), the acoustic signal simultaneously carries multiple types of information: (1) linguistic information (e.g. phonemic contrasts, lexical meaning), (2) non-linguistic information (i.e. relatively stable speaker characteristics, such as age or gender), and (3) paralinguistic information (i.e. dynamic aspects of the speaker, such as attitude and emotion). While this tripartite categorization is analytically useful, the acoustic signal does not neatly partition into these three separate streams—rather, linguistic and paralinguistic cues are almost always encoded in overlapping acoustic dimensions (cf. Scherer 1986, 2003). One important question in this research is thus to what extent paralinguistic information, such as emotional state, is encoded in the acoustic signal and how it is conveyed to the listener.

To bear on this general question, as a specific case study, we zoomed in on the auditory perception of "smiling" in Japanese—addressing whether "smiling" can be conveyed to the listeners without visual cues.[2] Our study builds upon a body of previous work exploring whether emotional states can be conveyed to listeners through the acoustic signal. Influential studies by Scherer (1986) and Banse and Scherer (1996), for example, demonstrated that different emotional states produce distinct acoustic profiles in speech, allowing listeners to identify emotions from vocal cues alone (see also Erickson 2005; Murray and Arnott 1993; Scherer 2003 for reviews). However, these studies focused primarily on prosodic features such as pitch and intensity as well as on general spectral characteristics (e.g. Kohler 2008; Nwokah 1999; Tartter 1980; Ruch 1993), leaving open the question of whether more fine-grained *segmental* features might also carry paralinguistic information. To that end, we made use of an articulatory phenomenon in Japanese that has so far received very little theoretical attention in the literature: the "labiodentalization" of bilabial consonants.

The traditional view in Japanese phonology is that /p/, /b/, and /m/ are, naturally, realized as bilabials, in which both lips are closed (Vance 1987, 2008). However, some scholars (e.g., Kawahara 2025; Kazama et al. 2004) have pointed out that in contexts involving a smile, these sounds can frequently be realized as labiodentals in which the upper teeth touch the lower lip—e.g. /b/ is produced as [v] rather than [b] and /m/ is produced as [ɱ], when the speaker is smiling. Impressionistically speaking, this use of labiodentals is especially commonly observed among young female speakers (Kazama et al. 2004). It is likely that this articulatory shift to labiodentals facilitates the maintenance of a smile, which would otherwise be interrupted by the lip closure required for a bilabial consonant; elevation of the lip corners in smiling presents an articulatory conflict with bilabial closure (Chan et al. 2018; Kang et al. 2021; Kawahara 2025). It may be the case that since bilabials and labiodentals are not phonemically contrastive

[2] Though see also Martin et al. (2017) on the functional roles of smiling in the context of social interactions, e.g. "being amicable," which does not necessarily involve actual "emotion."

in Japanese (Vance 1987, 2008), the language offers this sort of flexibility.[3]

From the speakers' perspective, this articulatory strategy can be understood as a means to continue smiles while producing labial sounds. However, what remains unclear is whether it possesses any psychological reality for the listener—in other words, do Japanese listeners use labiodentalization as a cue for smiling (see Aubergé and Cathiard 2003; Drahota et al. 2008; Tartter 1980; Tartter and Braun 1994 for previous studies on the perception of smiling)? If listeners have internalized a fine-grained phonetic association between labiodentals and smiling, we predict that labiodentalized stimuli will be judged as "smiling" more often than those tokens without labiodentalization. We also addressed the question of how the effects of labiodentals in Japanese, if any, would interact with the general effects of smiles on the voice quality, such as higher F0 and intensity (cf. Lasarcyk and Trouvain 1998).

In a nutshell, we addressed the question of whether smiles can be conveyed via speech signals without visual cues, focusing especially on the role of labiodentalized consonants in Japanese. If listeners can infer smiling gestures from purely auditory signals, this would constitute an interesting case of cross-modal perception—listeners are able to perceive information about one modality (i.e. visual facial expression) through cues presented in another modality (i.e. auditory speech signal). While cross-modal perceptual integration has been extensively documented in speech perception research (e.g. the McGurk effect: McGurk and MacDonald 1976 as well as Massaro and Cohen 1983 and Stiles et al. 2018; see Rosenblum 2008 and Spence 2011 for reviews), these phenomena typically involve the presentation of concurrent visual and auditory information. The case that we study in this paper is somewhat different, addressing whether listeners would be recovering absent visual information (i.e. the smile) from auditory cues alone, based on a learned association between a specific articulatory pattern and a facial gesture.[4]

---

[3] While this labiodentalization during smiled speech is an understudied phenomenon, it is documented in English (Chan et al. 2018) and Korean (Kang et al. 2021). It is interesting that this labiodentalization occurs even in a language in which [p] and [f] are contrastive (i.e. English). More cross-linguistic exploration is warranted to explore how wide-spread this pattern is across natural languages, and how the contrastive status between the two categories may or may not affect this labiodentalization. Also desired is a thorough quantitative sociophonetic analysis addressing how often this labiodentalization occurs by what kinds of speakers in what kinds of sociolinguistic contexts. Such exploration is warranted in Japanese and other languages.

[4] More distantly related may be the idea advanced by Ohala (1996) that we generally smile with the "[i]-face," since [i]'s high resonant frequency conveys the image of smallness via what is now well-known as the "frequency code".

Before we move on to the exposition of our experiments, we would like to share one personal story that partially motivated this research. The current experiments were inspired by a conversation that the first author had with a Japanese nurse during the COVID-19 pandemic. She reported that as a healthcare worker, she faced the challenge of conveying warmth to patients while wearing face masks, and felt that labiodental articulation helped signal a "smiling" voice even when facial expressions were obscured (see Martin et al. 2017 for this sort of interactive function of smiles). She specifically noted that the Japanese phrase [daidʒoːbɯ] 'it is okay/you are fine/no worries', very frequently used in the medical settings, contains [b], which can be pronounced as [v]. This observation prompted us to investigate systematically whether labiodentals can indeed function as perceptual cues for smiling in the absence of visual information, ultimately addressing the question of whether the nurse's smile was—and is—indeed conveyed to the patients over the mask.

2 Experiment 1

2.1 Method

To test the effects of general voice quality (the general acoustic consequence of smiling) and segmental articulation (the use of labiodentals), we conducted a perception experiment with these two factors fully crossed, resulting in four different conditions.

2.1.1 Stimuli

We selected 18 Japanese words beginning with bilabial consonants (/p/, /b/, /m/). To ensure the results were not an artifact of a specific lexical stratum, we included three word types: interjections, ideophones, and ordinary, prosaic words.[5] The full list of the stimuli is provided in Table 1.

[5] Interjections signal the speaker's emotional state or attitude (Ameka and Wilkins 2006), and previous work suggests that ideophones—imitative sensory words—also tend to be emotionally loaded (Baba 2003). Based on this literature, we expected that these two word classes might encode emotional parameters, such as smiling, more strongly than prosaic words. However, this expectation was largely speculative, and, as we will see below, there were no substantial effects of word types.

Table 1: The full list of the stimuli used in the experiment.

	Prosaic words	Ideophones	Interjections
C1 /m/	[mago] 'grandchild'	[mosaʔ] 'one's hair overgrown'	[maː] (admiration, surprise)
	[mɯsɯko] 'son'	[mɯkɯʔ] 'sitting up suddenly'	[moː] (annoyance)
C1 /b/	[bɯta] 'pig'	[bokoʔ] 'bumpy'	[beːʔ] (ridicule)
	[boːzɯ] 'close-cropped head'	[bɯɾaʔ] 'strolling'	[bɯːʔ] (disapproval)
C1 /p/	[pɯɾo] 'professional'	[pataʔ] 'falling flat'	[pɯɴ] (anger)
	[pasɯta] 'pasta'	[poɾoʔ] 'dropping'	[pɯʔ] (bursting into laughter)

2.1.2 Speakers and Recording

In order to obtain the speech stimuli for the current experiment, we asked two professional voice actors (both female—Speaker A and Speaker M) for the recording, because we suspected that non-professionals may not be able to manipulate smiles and the use of labiodentals in a thoroughly independent way. We specifically asked these two voice actors, because they use labiodentals in their own daily speech when they smile.

The voice actresses were instructed to produce each target word under four crossed conditions:

1. **Smile / Labiodental:** Speaking with a smile, using a labiodental articulation.
2. **Smile / Bilabial:** Speaking with a smile, maintaining a standard bilabial closure.
3. **Non-Smile / Labiodental:** Speaking with a neutral face, using a labiodental articulation.
4. **Non-Smile / Bilabial:** Speaking with a neutral face and standard bilabial closure.

This design yielded a total of 72 stimuli (3 initial sounds * 3 word types * 2 items per type * 4 conditions). The recorded tokens are available at the following OSF repository (https://osf.io/jhqw3/overview?view_only=14ab5584c22542f9bea91817e22ffc08).

2.1.3 Procedure

An online perception experiment was conducted via CrowdWorks (https://crowdworks.jp/), an online platform for crowd-sourcing service in Japan. A total of 200 participants, who were all native speakers of Japanese, were presented with the audio stimuli one by one in a randomized order, prepared in a Google Form. Their task was a binary forced-choice judgment: "Is the speaker smiling?" (Yes/No). No visual information was provided; judgment was based solely on the auditory signal.

To ensure data quality, we included two catch trials (an unambiguous laugh and an unambiguous sigh). The data from five participants who failed these check trials were excluded, leaving 195 valid participants for further analysis.

2.3 Results

Figure 1 summarizes the results of the perception experiment. The y-axis represents the cumulative proportion of listener responses, where the gray portion indicates a judgment of "smiling" and the white portion indicates "non-smiling." The data are faceted by the two experimental conditions: Voice Quality (Smile vs. Non-Smile) and Articulation (Bilabial vs. Labiodental), for each type of word type.

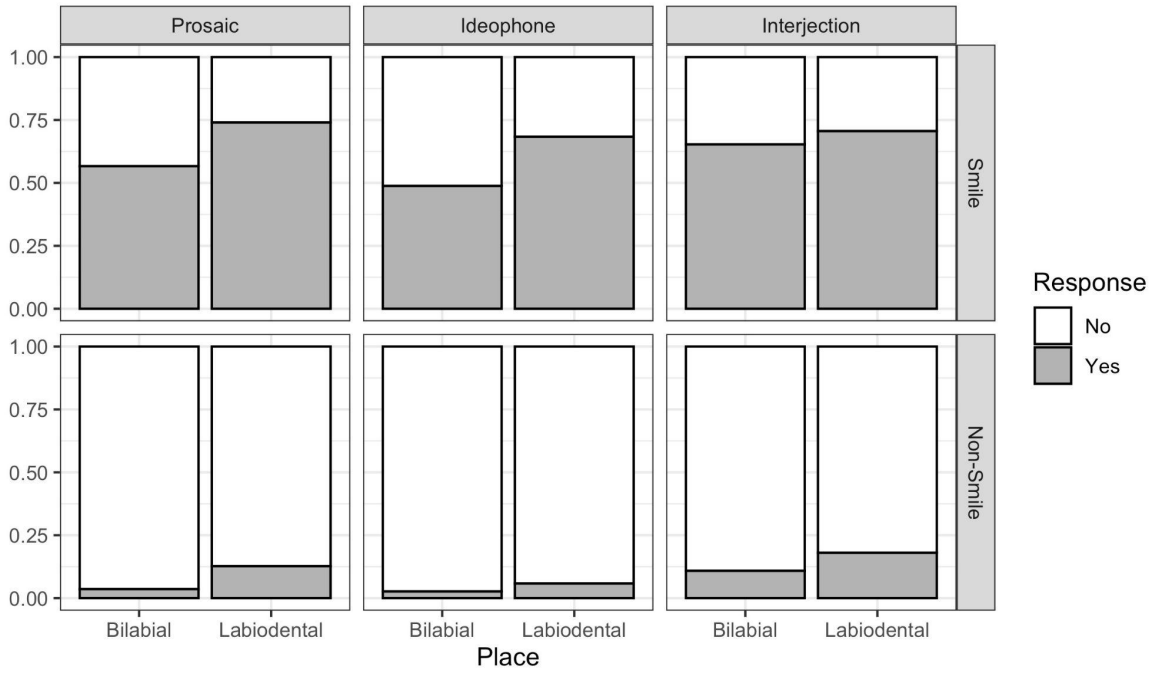

Figure 1: The results of Experiment 1. The y-axis represents the proportion of listener

judgments. Gray areas indicate the proportion of "smiling" responses, while white areas indicate "non-smiling" responses. The responses are broken down by Voice Quality (Smile vs. Non-Smile) and Articulation (Bilabial vs. Labiodental) for each lexical type.

As we can observe in Figure 1, the Voice Quality (the top three panels vs. the bottom three panels) had a substantial impact on the listeners' judgments. When the speakers were physically smiling (the bars at the top), at least more than 50% of the stimuli were correctly perceived as "smiling." On the other hand, when the speakers maintained a neutral expression (the bars on the bottom), the vast majority response was "non-smiling." This comparison confirms that the global acoustic modifications by actual smiles serve as a very robust perceptual cue for smiling.

More crucially, the results also reveal a substantial effect of Articulation (bilabials vs. labiodentals) as well. Both in the Smile and Non-Smile conditions, the stimuli produced with a labiodental articulation (the right bars within each facet) elicited a higher proportion of "smiling" responses compared to the bilabial baseline (the left bars within a facet). Finally, there does not seem to be any substantial effects of word type (the three different columns), which means that the effects of Voice Quality and Articulation hold robustly across these different lexical types.

To statistically assess the robustness of these factors, we analyzed the responses with a Bayesian mixed-effects logistic regression model using the brms package (Bürkner 2017) in R version 4.4.0 (R Core Team 2024). The dependent variable was the listener's judgment (smiling = 1, non-smiling = 0). The fixed effects were Voice Quality (Smile vs. Non-Smile), Articulation (Labiodental vs. Bilabial), word type (prosaic words as the baseline) and the two speakers. We also included random intercepts for participants and items; we did not include random slopes, because those models with random slopes did not converge. We employed weakly regularizing informative priors, ensuring that the posterior distributions were driven primarily by the data while preventing overfitting. Specifically, a Cauchy distribution (mean = 0, scale = 2.5) was assigned to the fixed effect coefficients, and a conservative Normal distribution (mean = 0, s.d. = 1) was assigned to the intercept (Gelman et al. 2008; Lamoine 2019).

The posterior distributions were estimated using four MCMC chains. Each chain consisted of 2,000 iterations, and the first 1,000 iterations were discarded as warm-up, yielding a total of 4,000 post-warmup samples for inference. The complete details of this statistical analysis, as well as the Bayesian posterior samples, are available at the OSF repository mentioned above.

Figure 2 illustrates the posterior distributions for the fixed effects in our Bayesian mixed-effects

logistic regression model. The circles represent the point estimate of the posterior distribution for each parameter. The horizontal bars represent the uncertainty around these estimates: the thick bars indicate the 80% credible intervals, while the thin extensions indicate the 95% credible intervals. Heuristically, we can interpret effects as "credible" when their 95% credible intervals do not include zero. Beyond this rule of thumb, however, Bayesian posterior distributions provide additional information: for each parameter, we can calculate the probability that the effect is positive or negative (denoted as $p(\beta > 0)$ or $p(\beta < 0)$), offering a more nuanced interpretation than binary significance hypothesis testing (see e.g. Kruschke 2014; Kruschke and Liddell 2018 for an accessible introduction to Bayesian analyses).

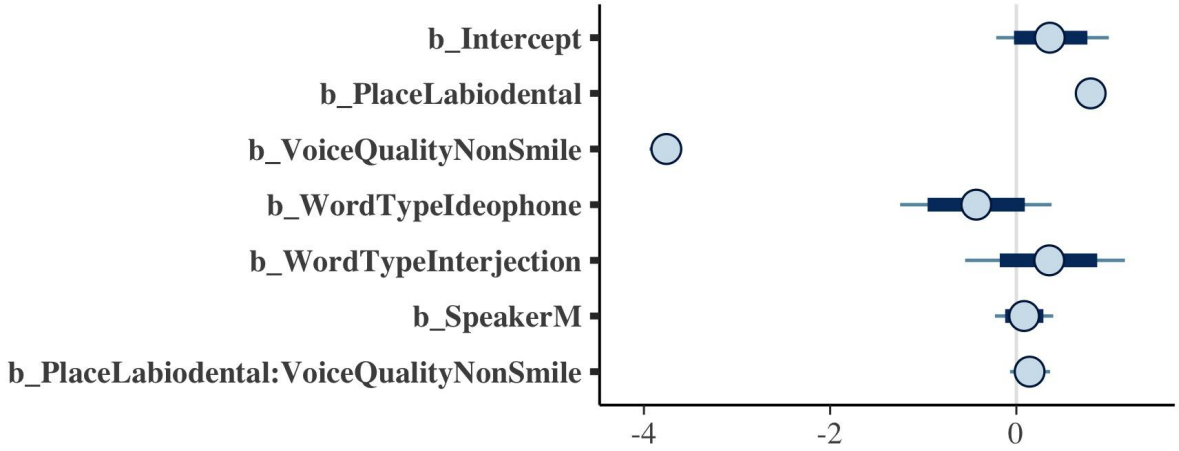

Figure 2: The MCMC plot of the posterior distributions of the fixed effects from the Bayesian regression model. The thick horizontal bars represent the 80% credible intervals, while the thin horizontal bars extend to the 95% credible intervals.

Starting with non-credible effects, none of the bottom four factors were credible (prosaic vs. ideophone; prosaic vs. interjection; Speaker A vs. Speaker M; the interaction between Articulation and Voice Quality). More interestingly, we observed a credible effect of Articulation (labiodentals, as opposed to bilabials). The 95% credible interval of this effect is [0.69, 0.91], with all the posterior samples being in the expected direction ($p(\beta > 0) = 1$). This result indicates that the stimuli containing labiodentals were more likely to be judged as "smiling", suggesting that the labiodental sound itself carries a "smiling" connotation.

The effect of Voice Quality (Smile vs. Non-Smile; the third factor from the top in Figure 2) was also very robust, with its 95% credible intervals being [−3.76, −3.47] and all the posterior samples being in the expected direction ($p(\beta < 0) = 1$). This result suggests that when the speakers are actually smiling, that acoustic signals contain enough information for the listeners to judge them as "smiling".

Finally, let us recall that we did not observe a strong interaction between these two main factors (the factor shown at the bottom in Figure 2), suggesting that the acoustic cue of the smile and the articulatory cue of the consonant function in an additive fashion.

2.4 Discussion

Experiment 1 tested whether a specific articulatory strategy—i.e. labiodentalization of bilabials—serves as a perceptual cue for smiling in Japanese. The results support the view that it does.

The experimental finding that Voice Quality substantially influenced the perception in the current study may not be too unsurprising; even over the phone, we can often tell whether the speakers are smiling or not (see Drahota et al. 2008; Tartter 1980; Tartter and Braun 1994 for previous studies on the perception of smiles; see also Erickson 2005 and Erickson et al. 2006 for the perception of other emotions). Nevertheless, we find it important that the current experiment confirmed this impressionistic observation in a systematic and quantitative way in Japanese.

Moreover, the finding that labiodentalization acts as an independent cue was surprising and, we believe, theoretically significant. It implies that Japanese listeners have learned a specific association—"if labiodental sounds, then smiling face"—and are able to use this association to infer whether the speakers are actually smiling or not. This association allows them to infer the speaker's facial expression *even when the speakers are actually not smiling* (e.g. in the Non-Smile/Labiodental condition in Figure 1).

We contend that this result implies that labiodental articulation has entered what is known as "indexical field" (Eckert 2008, 2012) of Japanese listeners—a constellation of (social) meanings that become associated with a linguistic variant. In this framework, phonetic variants can acquire social meanings that extend beyond their original phonetic or physiological origins. In our case, while labiodentalization is likely to have been originated as a physiological byproduct of smiling (it is easier to maintain a smile without full lip closure: Chang et al. 1998; Kang et al. 2021; Kawahara 2025), it appears to have taken on independent social meaning for Japanese listeners, indexing a 'smiling voice' even when the speaker is not actually smiling.

3 Experiment 2

3.1 Introduction

While the results of Experiment 1 revealed a robust effect of labiodentals, it may have been the case that those labio-dental tokens, regardless of whether they are in the Smile or Non-Smile condition, were produced with higher F0 and higher intensity (see Lasarcyk and Trouvain 2008 for higher F0 during smiled speech). Some of us had this auditory impression as trained linguists, and felt it necessary to address this concern. First, as a post-hoc acoustic analysis, we measured the average F0 of each token during their first vowels, by creating a 20 ms analysis window aligned at the center of the vowel; recall that the target consonants (bilabial vs. labiodental) were in the first syllables (Table 1). The extraction process was automated using Praat. The results of this post-hoc acoustic analysis appear in Figure 3:

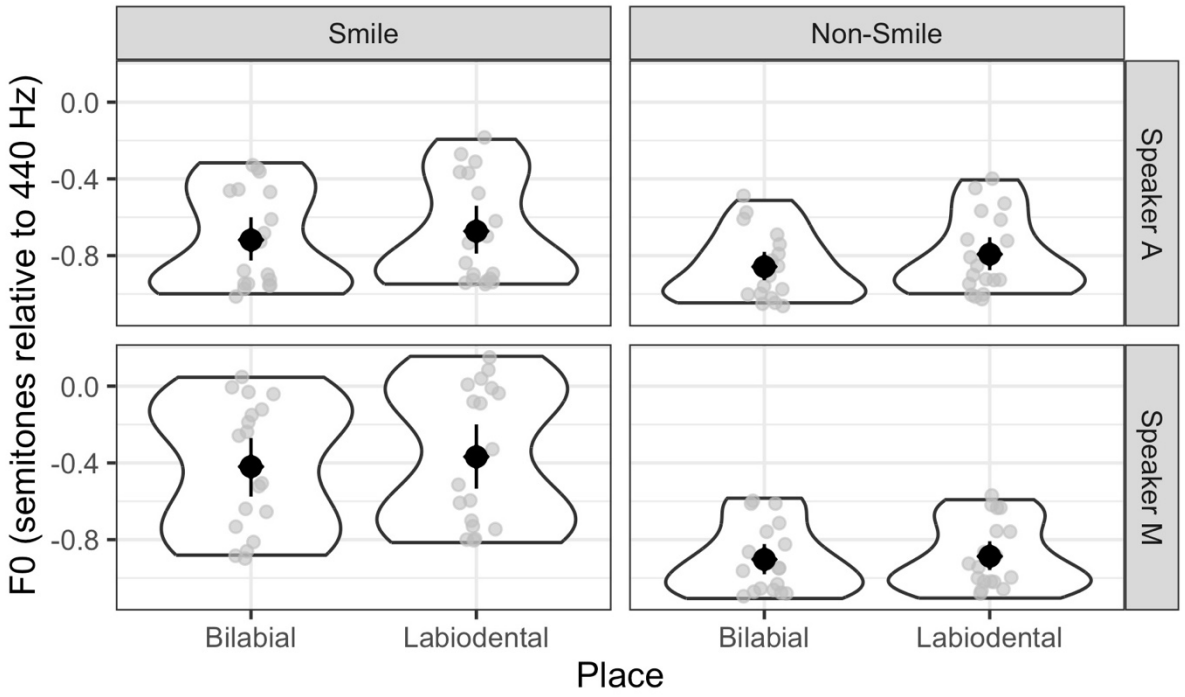

Figure 3: The results of the post-hoc acoustic analysis of the stimuli used in Experiment 1. The y-axis represents the average fundamental frequency (F0) of the first vowel, expressed in semitones relative to 440 Hz. The error bars represent the 95% confidence intervals. The data are faceted by Voice Quality (Smile vs. Non-Smile) and Speaker (Speaker A vs. Speaker M).

We observe that generally speaking, the labiodentals (the right bars within each facet) have higher F0 compared to the bilabials (the left bars within each facet), although this difference may not be very substantial in the Non-Smile condition for Speaker M (the right bottom facet). This general pattern raises the possibility that the effects of the labiodentals may have been confounded by this effect of F0 in Experiment 1. We addressed this concern in Experiment 2.

3.2 Method

In order to control for the effects of F0, different renditions of the same word produced by a single speaker were first aligned in pitch and then in time with Vocalign Pro (Synchro Arts by LANDR, v. 6.1.30), using as reference non-smiling renditions featuring bilabial consonants. The audio plugin was set to match both pitch and time with the highest precision and resolution. In addition, the aligned recordings were RMS-equalized to the same reference, trimming the silence around an utterance. This equalization was facilitated with a self-developed script in Python. Silence was considered to be 30 dB below the maximum RMS value of the signal. The edited stimuli can be examined at the OSF repository.

The methodological details for the perceptual experiment were almost identical to those of Experiment 1. A total of 201 native speakers of Japanese were recruited for the experiment. After excluding four participants who failed the check trials, the data from 197 participants remained for further analysis. The full details of the Bayesian analysis as well as the posterior samples are available at the OSF repository.

3.3 Results

The results of Experiment 2 appear in Figure 4. As with Experiment 1, the y-axis represents the proportion of listener judgments (gray = “smiling”; white = “non-smiling”). The responses are organized by the two conditions: Voice Quality (Non-Smile vs. Smile) and Articulation (Bilabial vs. Labiodental). The crucial comparison between bilabials and labiodentals appears within each facet.

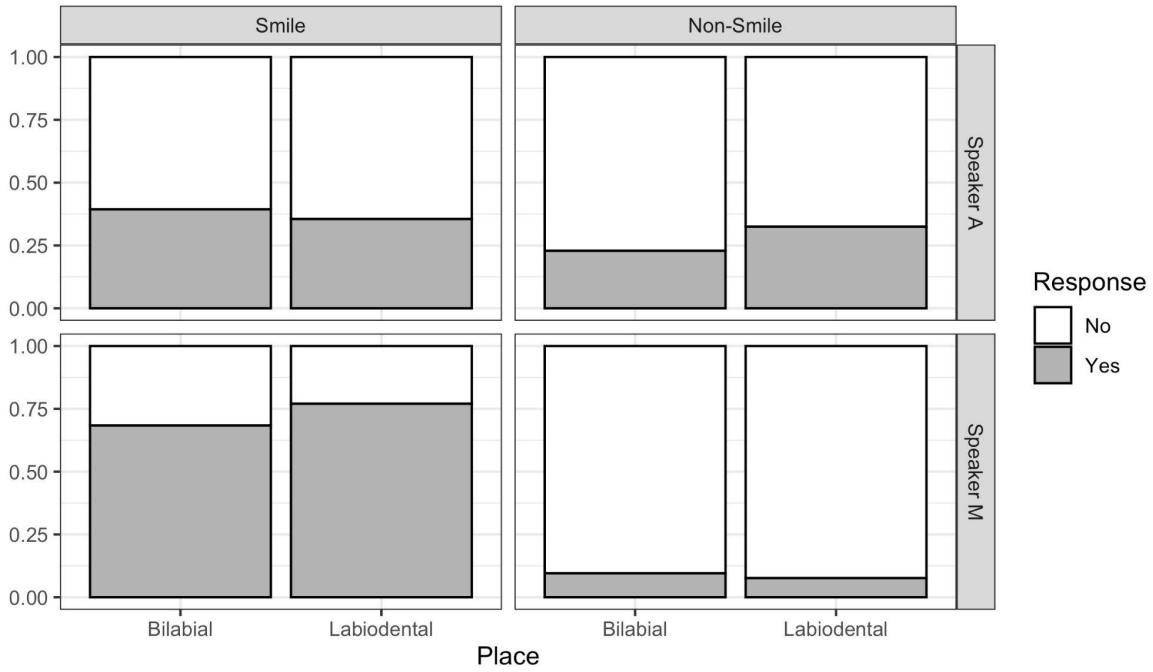

Figure 4: The results of Experiment 2 (acoustically controlled stimuli). The y-axis represents the proportion of listener judgments (gray = "smiling"; white = "non-smiling"). The results are organized by Voice Quality (Non-Smile vs. Smile) and Articulation (Bilabial vs. Labiodental). Within each facet, the right bar represents the Labiodental condition and the left bar represents the Bilabial condition. Unlike Experiment 1, this experiment used only prosaic words.

The results are not as straightforward as those of Experiment 1, but we observe some interesting patterns. First, we observe that voices in the smile condition generally induced more "smile responses" (the left vs. right column), even in the current experiment, in which the F0 and amplitude are controlled; this result is likely because smiling may influence the general formant frequency structure, due to various articulatory factors, such as spread lips, shortened vocal tract, and possibly raised larynx (e.g. Kohler 2008; Nwokah 1999; Tartter 1980; Ruch 1993).

Zooming in on the effect of labiodentals, for Speaker A, labiodentals induced more smiling responses compared to the bilabials, but only for those speech that were produced without smiles (the top right facet). The opposite pattern holds for Speaker M: labiodentals induced more smiling responses for those stimuli that were originally produced with smiles (the bottom left facet).

Figure 5 illustrates the results of the Bayesian mixed-effects model. The effect of Articulation (labiodentals as opposed to bilabials) was almost credible, with its 95% credible interval being [−0.4, 0.07]. The posterior probability of this coefficient being negative was 0.94. The effect of Voice Quality (whether the speakers were actually smiling or not) was very robust this time as well, with its 95% credible interval being [−1.20, −0.65], with all the posterior samples being

negative. However, as we observed in Figure 4, the picture is complicated so that the three-way interaction between Smiling, Articulation and Speaker was credible, with its 95% credible interval being [−2.17, −0.89].

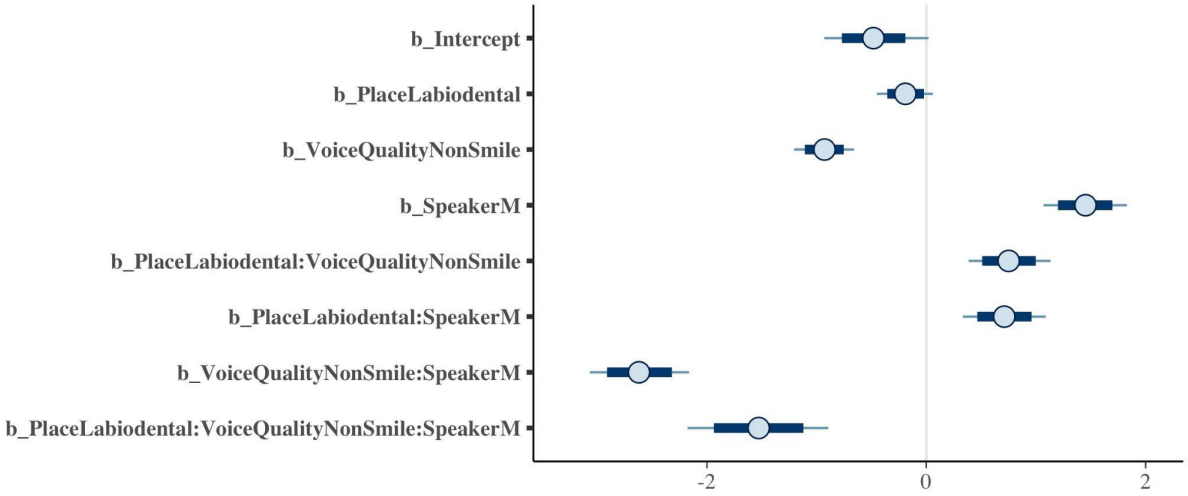

Figure 5: Posterior distributions of the fixed effects from the Bayesian mixed-effects logistic regression model for Experiment 2.

3.4 Discussion

To summarize the main finding of this experiment, we first of all found that even when the F0 contour and amplitude are fully controlled, there is a substantial effect of Smile vs. Non-Smile in Japanese. This finding, we think, is important, as it shows that there is more than F0 and amplitude (e.g. general spectral characteristics), which is associated with smiled speech. Further, the effects of labiodentality were observed in the current experiment as well, but how they manifest themselves depends on the speaker: for Speaker A, it increased the smiling responses in the Non-Smile condition whereas for Speaker M, it increased the smiling responses in the Smile condition.

We are unable to offer a definitive answer as to why this complicated pattern holds, but several possibilities warrant consideration. The first possibility is that the acoustic manipulation (F0 and amplitude normalization) may have inadvertently created unnatural stimuli that disrupted normal perceptual processes differently for the two speakers.

The second possibility would be that different speakers may employ different strategies for coordinating labiodentalization with other smile-related acoustic cues—for Speaker A, labiodentals may be more diagnostic in the absence of other smile cues, while for Speaker M, they may serve as secondary “enhancement” of existing, primary smile cues, so that labiodentals

show a tangible effect only when other primary cues are available (cf. Stevens and Keyser 1989).

Since there are multiple ways in which smiles can influence articulatory settings (e.g. shortening of the vocal tract and raising of the larynx), it is possible that these two speakers differ in terms of how they articulatorily implement their smiled speech. This hypothesis needs to be explored in a separate experiment which simultaneously record their articulation and acoustics, ideally with more than two speakers.

Despite these interpretive challenges, however, Experiment 2 confirmed that labiodentals can play a role in conveying smiles independently of F0 and amplitude cues, albeit in ways that may be more complex and speaker-dependent than Experiment 1 suggested.

4 General Conclusion

In conclusion, while we often assume that we "hear a smile" primarily through global prosodic cues—such as raised F0 and spectral characteristics—our experiments suggest that we may also, in a literal sense, "hear the lips." The results of Experiment 1 demonstrated that Japanese listeners perceive labiodental articulation as a reliable marker of smiling, even when the speaker is not actually smiling (cf. Drahota et al. 2008; Erickson 2005; Erickson et al. 2006; Tartter 1980; Tartter and Braun 1994). Experiment 2 further solidified this finding by showing that this perceptual effect persists even when fundamental frequency, utterance timing, and amplitude are strictly controlled (although this effect was tangible in different contexts across the two different speakers). Together, these findings indicate that Japanese speakers utilize fine-grained segmental cues—specifically the shift from bilabial to labiodental—to infer the non-visible facial gestures of speakers.

Returning to the personal story that motivated this research, our studies provide (albeit indirect) evidence that when nurses attempt to convey their smiles over masks—whether through labiodental articulation or general prosodic manipulations—those smiles can indeed be conveyed to their patients. Indeed, the very fact that the nurse reported consciously employing labiodental articulation as a communicative strategy suggests that this phenomenon has moved beyond a mere mechanical byproduct of smiling (cf. Chang et al. 2018; Kang et al. 2021); it appears to have been conventionalized—and arguably "phonologized"—as an independently deployable resource for signaling warmth.

The current results we believe have broader implications for the architecture of speech perception. The traditional view often compartmentalizes speech processing into linguistic decoding (identifying phonemes and words) and paralinguistic processing (identifying emotion and

attitude) (e.g. Abercrombie 1967; Fujisaki 2004; Laver 1994, though cf. Foulkes and Docherty 2006). However, our finding that a specific allophonic variant ([v] for /b/, [m̪] for /m/, etc.) can serve as a cue for an emotional state challenges a strict modular separation. Allophonic variants are not only conditioned by phonetic/phonological factors, but also by a paralinguistic factor (i.e. smiling), and this variation can serve as a reliable perceptual cue to deliver that gesture.

We acknowledge, however, that the scope of the current study necessitates some caution regarding the generalizability of the results. Specifically, we relied on stimuli produced by professional voice actors. Recall that this design choice was deliberate (and even necessary); we reasoned that it was very likely that untrained speakers were going to struggle to dissociate the articulatory gesture (labiodentalization) from the global physiological state (smiling) in the independent manner required for our experimental design. However, this choice of ours raises a question regarding whether this phenomenon is limited to a "performative" register used by professionals, or reflects a broader communicative reality in daily life.

Future research should therefore address whether our main findings generalize to tokens produced by lay speakers in naturalistic settings. It may find interesting differences between voice professionals and lay people; it may additionally be also informative to compare actors and voice actors, since voice actors are, after all, professionals who convey their emotional status only through their voice and thus may be better at conveying emotional states through their voice.

Several other specific questions emerge for future research. First, is labiodentalization indeed more prevalent among certain demographic groups (e.g. younger speakers, female speakers), as hinted by Kazama et al. (2004)? If so, do listeners from different demographic groups show different sensitivity to this cue? Does the strength of the labiodental–smile association vary across different speaking styles or registers? Finally, comparative work examining whether similar phenomena exist in other languages would help determine whether this represents a language-specific convention or a more universal strategy for maintaining smiles during speech (Chang et al. 2018; Kang et al. 2021). Our experiments open up opportunities for these new research questions.

Overall, this study highlights the rich, multimodal nature of speech communication. Even in the absence of visual information, listeners are surprisingly adept at recovering the physical state of the speaker's face, provided the phonological grammar of their language offers the necessary clues.

References

Abercrombie, David. 1967. *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Ameka, Felix K. & David P. Wilkins. 2006. Interjections. In Jan-Ola Östman & Jef Verschueren (eds.), *Handbook of Pragmatics*, 1–22. Amsterdam: John Benjamins.

Aubergé, Véronique & Marie-Agnès Cathiard. 2003. Can we hear the prosody of smile? *Speech Communication* 40(1-2). 87–97.

Baba, Junko. 2003. Pragmatic function of Japanese mimetics in the spoken discourse of varying emotive intensity levels. *Journal of Pragmatics* 35(12). 1861–1889.

Banse, Rainer & Klaus R. Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology* 70(3). 614–636.

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–28.

Chan, Terrina, Ryan C. Taylor, Esther Y. Wong & Bryan Gick. 2018. Coarticulation of speech and smile movements. *The Journal of the Acoustical Society of America* 144. 1903–1904.

Drahota, Amy, Alan Costall & Vasudevi Reddy. 2008. The vocal communication of different kinds of smile. *Speech Communication* 50(4). 278–287.

Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4). 453–476.

Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41. 87–100.

Erickson, Donna. 2005. Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology* 26(4). 317–325.

Erickson, Donna, Kenji Yoshida, Caroline Menezes, Akinori Fujino, Takemi Mochida & Yoshiho Shibuya. 2006. Exploratory study of some acoustic and articulatory characteristics of sad speech. *Phonetica*, 63(1), 1-25.

Foulkes, Paul & Gerard J. Docherty. 2006. The social life of phonetics and phonology. *Journal of Phonetics* 34(4). 409–438.

Fujisaki, Hiroya. 2004. Information, prosody, and modeling — with emphasis on tonal features of speech. In *Proceedings of Speech Prosody 2004*, 1–10. Nara, Japan.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau & Yu-Sung Su. 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2(4). 1360–1383.

Johnson, Keith. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34(4). 485–499.

Kang Elisabeth H., Yadong Liu, Annabelle Purnomo, Melissa Wang & Bryan Gick. 2021. Speaking versus smiling: The labiodentalization of bilabials in Korean. *Canadian Acoustics* 49(3). 36–37.

Kawahara, Shigeto. 2025. *Koe no gengogaku nyūmon* [An introduction to the linguistics of "voice"]. Tokyo: NHK Publishing.

Kazama, Kiyomi, Zendo Uwano, Kazuto Matsumura & Ken Machida. 2004. *Gengogaku* [Linguistics] (2nd edn.). Tokyo: University of Tokyo Press.

Kohler, Klaus J. 2008. 'Speech-smile', 'speech-laugh', 'laughter' and their sequencing in dialogic interaction. *Phonetica* 65(1–2): 1–18.

Kruschke, John K. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd edn.). Amsterdam: Academic Press.

Kruschke, John K. & Torrin M. Liddell. 2018. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review* 25(1). 178–206.

Lasarcyk, Eva & Jürgen Trouvain. 2008. Spread lips + raised larynx + higher F0 = smiled speech?: An articulatory synthesis approach. *Proceedings of the 8th International Seminar on Speech Production*, 345–348.

Laver, John. 1994. *Principles of phonetics*. Cambridge: Cambridge University Press.

Lemoine, Nathan P. 2019. Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos* 128(7). 912–928.

Martin, Jared, Magdalena Rychlowska, Adrienne Wood & Paula Niedenthal. 2017. Smiles as multipurpose social signals. *Trends in Cognitive Science* 21(11). 864–877.
Massaro, Dominic W. & Michael M. Cohen. 1983. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 9(5). 753–771.
McGurk, Harry & John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264(5588). 746–748.
Murray, Iain R. & John L. Arnott. 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America* 93(2). 1097–1108.
Nwokah, Eva E., Hui-Chin Hsu, Patricia Davies & Alan Fogel. 1999. The integration of laughter and speech in vocal communication: A dynamic systems perspective. *Journal of Speech, Language, and Hearing Research* 42. 880–894.
Ohala, John J. 1996. Ethological theory and the expression of emotion in the voice. In *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP 96)*, vol. 3, 1812–1815. Philadelphia, PA.
R Core Team. 2024. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
Rosenblum, Lawrence D. 2008. Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science* 17(6). 405–409.
Ruch, Willibald. 1993. Exhilaration and humor. In Lisa Feldman Barrett, Michael Lewis & Jeannette M. Haviland-Jones (eds.), *The handbook of emotions*, 605–616. New York: The Guilford Press.
Scherer, Klaus R. 1986. Vocal affect expression: A review and a model for future research. *Psychological Bulletin* 99(2). 143–165.
Scherer, Klaus R. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40(1-2). 227–256.
Spence, Charles. 2011. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics* 73(4). 971–995.
Stevens, Kenneth N. & Samuel Jay Keyser. 1989. Primary features and their enhancement in consonants. *Language* 65(1). 81–106.
Stiles, Noelle R. B., Monica Li, Carmel A. Levitan, Yukiyasu Kamitani & Shinsuke Shimojo. 2018. What you saw is what you will hear: Two new illusions with audiovisual postdictive effects. *PLoS ONE* 13(10). e0204217.
Tartter, Vivien C. 1980. Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics* 27(1). 24–27.
Tartter, Vivien C. & Diane Braun. 1994. Hearing smiles and frowns in normal and whisper registers. *The Journal of the Acoustical Society of America* 96(4). 2101–2107.
Vance, Timothy J. 1987. *An introduction to Japanese phonology*. Albany: State University of New York Press.
Vance, Timothy J. 2008. *The sounds of Japanese*. Cambridge: Cambridge University Press.