Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 917–922

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Audible smiles and frowns affect speech comprehension

Hugo Quené [a,*], Gün R. Semin [b], Francesco Foroni [b]

[a] Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands
[b] Faculty of Social and Behavioral Sciences, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, The Netherlands

Received 15 November 2011; received in revised form 8 February 2012; accepted 12 March 2012
Available online 20 March 2012

Abstract

Motor resonance processes are involved both in language comprehension and in affect perception. Therefore we predict that listeners understand spoken affective words slower, if the phonetic form of a word is incongruent with its affective meaning. A language comprehension study involving an interference paradigm confirmed this prediction. This interference suggests that affective phonetic cues contribute to language comprehension. A perceived smile or frown affects the listener, and hearing an incongruent smile or frown impedes our comprehension of spoken words.

Keywords: Smiles; Speech comprehension; Emotion; Affect perception; Motor resonance

1. Introduction

In spoken language, vowels and consonants convey the linguistic meaning intended by the speaker. In addition, and unlike written language, speech also conveys a speaker's emotional state (Williams and Stevens, 1972; Frick, 1985; Neumann and Strack, 2000; Scherer, 2003; Batliner et al., 2003) mainly by means of its prosody. In addition, the audible properties of the speaker's vocal tract, in particular its second spectral resonance (formant F2) as well as the dispersion between formants, convey whether or not a speaker is smiling while talking (Ohala, 1980, 1983). Emotionally and affectively nuanced utterances play a central role in speech communication, by conveying importance, relevance, urgency, and attitude, in addition to the spoken semantic content. Listeners can decode audible affective cues such as smiles and frowns (Tartter and Braun, 1994), even with unfamiliar speakers (Drahota et al., 2008) and in foreign languages (Pell et al., 2009).

In this study we hypothesize that comprehension of a word's semantic meaning and affect perception based on its phonetic form, are not separate, but interacting components of spoken word processing. The presumed causal mechanism for this interaction is motor resonance (Gallese et al., 1996; Kohler et al., 2002) which is involved in listeners' retrieval of linguistic meaning (Wilson et al., 2004; Zwaan and Taylor, 2006), as well as in perception of affect (Gallese, 2003, 2009; Niedenthal, 2007; Foroni and Semin, 2009). Thus we investigate whether affectively meaningful phonetic features, related to affective facial expressions such as smiles and frowns, also influence spoken word recognition. We predict that spoken word perception will be faster if the semantic meaning and the affective phonetic cues of a word are congruent, relative to spoken words with incongruency between semantic content and affective phonetic form. This incongruence would yield a phonetic Stroop effect (Stroop, 1935): if the positive word *pleasant* is spoken with an incongruent affective phonetic form (i.e., frown), its semantic evaluation is predicted to be slower than if this positive word *pleasant* is spoken with a congruent smile.

Previous studies have already shown that emotionally and socially incongruent phonetic forms do indeed have a

* Corresponding author. Tel.: +31 30 2536070; fax: +31 30 2536000.
E-mail addresses: h.quene@uu.nl (H. Quené), g.r.semin@uu.nl (G.R. Semin), f.foroni@uu.nl (F. Foroni).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.03.004

negative effect on speech processing. For example, semantic evaluation of happy, neutral and angry words was found to be slower if these emotional words were spoken with incongruent emotional prosody (Schirmer and Kotz, 2003; Mehrabian and Wiener, 1967; Grimshaw, 1998; Schirmer et al., 2002). Similarly, naming latencies for happy, neutral and sad words were longer if the emotional words were spoken with incongruent emotional prosody (Nygaard and Queen, 2008). In an eye-tracking study with visually presented faces expressing various emotions, listeners gazed more frequently to faces with emotions congruent to the prosody of the speech stimuli (Paulmann et al., 2012). Spoken sentence comprehension was also affected (as indicated by an N400 effect) by inconsistency or incongruence between the semantic content and the speaker characteristics (gender, age, and social status) expressed by the speaker's voice (Van Berkum et al., 2007; Tesink et al., 2008).

The present study aims to expand this converging evidence, in three ways. First, our focus is on smiles and frowns as affective facial gestures, and not on phonetic expressions of basic emotions (Scherer, 2003). Because smile gestures and frown gestures necessarily interfere with speech production, the perception of such affective speech may show stronger motor resonance (Gallese et al., 1996; Kohler et al., 2002; Wilson et al., 2004; Zwaan and Taylor, 2006; Gallese, 2003, 2009). One drawback is that the affective meanings of smiles and frowns may be ambiguous. A smile, for example, might express enjoyment, friendliness, and/or dominance (Niedenthal et al., 2010).

For similar reasons, secondly, our focus is not on prosody (e.g. Scherer, 2003; Nygaard and Queen, 2008; Schröder, 2006) but on formant frequencies (and hence formant dispersion) as the main auditory cue. In case of a human speaker, the pattern of formant frequencies may be regarded as the audible effect of a smile or frown gesture produced simultaneously with the speech (Ohala, 1980; Tartter and Braun, 1994; Chuenwattanapranithi et al., 2008; Lasarcyk and Trouvain, 2008). Other effects of smiles and frowns, mainly expressed prosodically by means of F0 (Ohala, 1980; Tartter and Braun, 1994; Chuenwattanapranithi et al., 2008; Lasarcyk and Trouvain, 2008), are ignored in this study, because these prosodic effects cannot be easily related to motor resonance processes. Although this limitation in phonetic cues may result in a conservative study, we note that smiles and frowns are also perceived in whispered speech without F0 (Tartter and Braun, 1994), and that spectral cues appear to be more important than F0 cues for perception of affect (Xu and Kelly, 2010).

Thirdly, the effects of affective incongruence are investigated here not by means of acted speech (e.g. Grimshaw, 1998; Schirmer et al., 2002; Nygaard and Queen, 2008; Paulmann et al., 2012) but by means of synthesized speech in which formants were manipulated (Chuenwattanapranithi et al., 2008; Lasarcyk and Trouvain, 2008). This phonetic simulation of smiling and frowning allows stronger experimental control over the affective phonetic cues contributing to spoken word processing. Thus a word's affective meaning and its affective phonetic form were varied orthogonally, yielding congruent and incongruent combinations of affective meaning and form. The listeners' task involved language comprehension of positively and negatively valenced words. Words are predicted to be understood slower if spoken in an incongruent form (e.g., positive words with frown) than in a congruent form (e.g., positive words with smile).

2. Method

2.1. Stimulus words

Experimental stimuli consisted of 60 Dutch words (30 having positive meaning, e.g. *eerlijk* "honest", and 30 having negative meaning, e.g. *vijandig* "hostile"). This selection was based on a pre-test in which words were rated for affective value by a sample of 30 Dutch students. Positive words were rated more positively ($M = 7.45$, $SD = 0.49$) than negative words [$M = 2.85$, $SD = 0.50$, $t(58) = 32.62$, $p < 0.001$] on a 9-point scale. Positive words had the same length in syllables ($M = 2.6$, $SD = 1.0$) as negative words [$M = 2.6$, $SD = 0.8$, $t(58) = 0.29$, $p = 0.774$]. A male native speaker read each word in an affectively neutral manner, using a randomized list of words, and reading each word as a separate utterance (without list intonation). These readings were recorded and then used as targets for speech synthesis.

2.2. Stimulus preparation and selection

Spectral resonances (formants) were computed from the neutral speech recordings, and checked manually before being used for speech synthesis. The corrected formant values were used to control a formant-based speech synthesizer (Klatt, 1980; Boersma and Weenink, 2011). For neutral phonetic forms, the unshifted frequencies of the formants were used. For smiling forms, the frequency of the lowest spectral resonance (formant F1) was shifted up by 5%, and frequencies of higher formants (F2 to F5) were shifted up by 10% (Ohala, 1980). Conversely, for frowning forms, the F1 was shifted down by 5%, and higher formants were shifted down by 10% (Schirmer et al., 2002). Formants were adjusted throughout the target word. This resulted in phonetically neutral synthetic realizations (positive–neutral, negative–neutral), or congruent realizations (positive–smiling, negative–frowning), or incongruent realizations (negative–smiling, positive–frowning). All other synthesis parameters were identical in corresponding neutral, congruent and incongruent forms of a word. The pitch contour was copied from the original recording.

2.3. Pre-tests

In order to verify the noticeability of the phonetic manipulations, as well as the resulting intelligibility of the