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 target words, two pre-tests were conducted. In the first pre-test, listeners rated each spoken word on a 9-point scale, to indicate to what extent the word was spoken with a simultaneous frown (scale value 1), in neutral fashion (value 4), or with a simultaneous smile (value 9). The three phonetic forms (unshifted or "neutral", with formants shifted down or "frowning", and with formants shifted up or "smiling") were counterbalanced over three lists. Each list was presented over loudspeakers to a separate group of listeners (of 42, 16, and 16 listeners, respectively), at an inter-stimulus interval of 2 s (as established in pilot tests). Listeners were native Dutch-speaking undergraduate students at Utrecht University, without any knowledge about the purpose of the study. They were asked to rate the presence of simultaneous affective gestures during speech production, on a scale with end points marked with a frowning face symbol (☹, scale value 1) and a smiling face (☺, value 9). These subjective ratings were analyzed by means of mixed-effects regression, with listeners and target words as crossed random effects (Baayen et al., 2008; Quené and Van den Bergh, 2008). In the fixed part of the model, the "frowning" and "smiling" forms were compared relative to the neutral (unshifted) phonetic form; semantic valence was also included as a fixed effect.

Subjective ratings were significantly lower for the "frowning" forms with formants shifted down (relative to "neutral" forms, $\beta = -1.54$ scale point, *s.e.* $= 0.07$, $p = 0.0001$), and significantly higher for the "smiling" forms with formants shifted up ($\beta = 1.00$ scale point, *s.e.* $= 0.07$, $p = 0.0001$), as illustrated by the boxplots in Fig 1. This confirms that the phonetic manipulations of formants in the speech stimuli do indeed successfully convey the desired affective property, viz. of speech being produced with a simultaneous frown gesture or smile gesture. Moreover, subjective ratings were also higher for semantically positive words than for semantically negative words ($\beta = 0.78$ scale point, *s.e.* $= 0.17$, $p = 0.0001$). Semantic valence thus yields a significant effect on subjective ratings

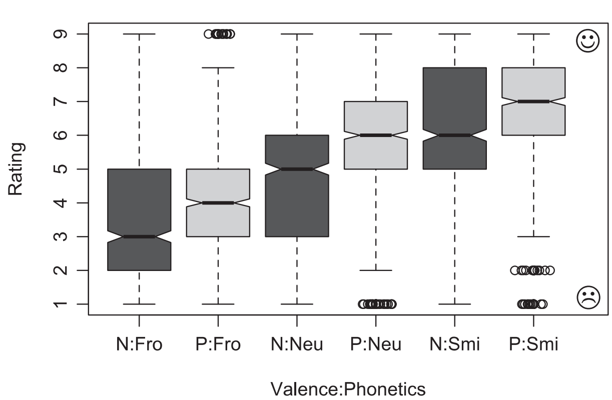

Fig. 1. Boxplots of subjective ratings as to what extent words were spoken with a simultaneous smile, in neutral fashion, or with a simultaneous frown, broken down by semantic valence (N: negative, darker boxes, P: positive, lighter boxes) and by phonetic form ("frowning", neutral, "smiling"). Notches indicate approximate 95% confidence intervals of the box median.

of a phonetic affective property. This interesting main effect supports our research hypothesis, as it suggests that subjects' task of affect perception at the phonetic level was not entirely separate from semantic evaluation, thus indicating interacting processes. No interaction effect was observed between phonetic form and semantic valence on subjective ratings: the effects of phonetic manipulations are the same for negative and for positive words.

In the second pre-test, the phonetically neutral forms of the 60 target words were presented to 15 listeners (from the same sample of participants as below) in randomized order, to assess intelligibility of the synthesized target words. Participants listened to each resynthesized word (with unshifted formants) individually, and typed the word they had heard. Responses were scored for accuracy, with correction of occasional spelling errors (e.g. "interesant" for *interessant*). Participants' typed responses showed poor intelligibility (accuracy < 0.8) for 7 out of 60 target words (5 positive, 2 negative). These 7 poorly intelligible target words were kept in the main experiment but were excluded from further data analysis.

2.4. Participants and procedure

In the main experiment, 48 native Dutch-speaking students (39 females, 9 males) with no hearing, language or speech deficits listened to the synthesized words. Participants' ages were between 18 and 27 years (median 21.5 years). The three phonetic forms of a word were balanced over three separate experimental lists. Each list was presented to 16 participants. Listeners' task was to classify the meaning of the spoken word as positive (exemplar *peace*) or negative (exemplar *war*) as quickly and as accurately as possible after the offset of the spoken word. Stimuli were presented with a 5-ms fade-in and fade-out to prevent click sounds. Before the actual task, participants were presented with 5 practice trials (including all 3 phonetic forms, both congruent and incongruent). The actual test started with 10 warm-up items indiscernible from the subsequent stimuli; the list of stimuli was re-randomized for each participant. Listeners responded by pressing one of two response buttons, always using the index finger of their dominant hand. Positive and negative response buttons were balanced over the 16 participants within each experimental list. No instructions were given about the relevant semantic or phonetic properties of the stimuli. The total time of an experimental session was about 12 min.

3. Results

The main dependent measure was response time (RT) measured from the onset of the spoken word. Responses with outlier RTs (4%) and incorrect responses (4%) were excluded from the data analysis. The remaining RTs were analyzed by means of mixed-effects regression, with listeners and target words as crossed random effects (Baayen et al., 2008; Quené and Van den Bergh, 2008). The resulting

Table 1
Estimated parameters of mixed-effects model of response times. Estimates of fixed parameters are given in ms, with standard error and significance level (in boldface if $p < 0.05$). Estimates of random parameters are given in standard deviations, with 95% confidence interval of the estimate. $N = 2343$.

Fixed coefficients	Estimate	s.e.	*p*
Intercept	1264.2	30.6	0.0001
Positive valence word	−119.7	33.5	**0.0002**
Incongruent phonetic form	39.4	13.8	**0.0036**
Congruent phonetic form	1.5	13.8	0.9170
Pos. valence × incongruent	−4.2	20.1	0.8336
Pos. valence × congruent	10.4	20.1	0.6028
Random coefficients	**Estimate**	**95% C.I.**	
Listeners	140.2	96.7, 136.6	
Target words	110.4	81.0, 115.9	
Residual	197.8	193.9, 205.4	

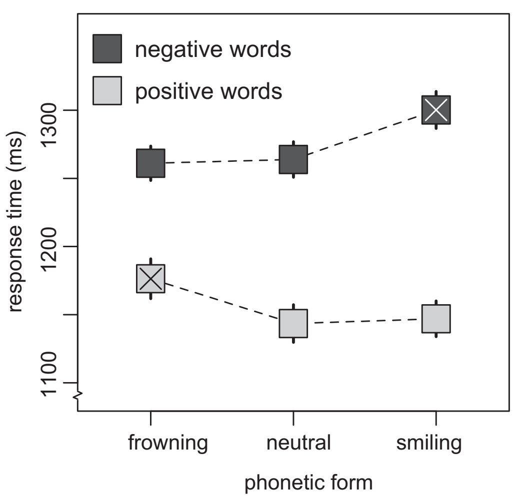

Fig. 2. Mean response times for negative and positive target words, broken down by phonetic form. Error bars indicate standard error of the mean. Dashed lines are for visual guidance only. Conditions with incongruent phonetic forms are crossmarked.

optimal model, summarized in Table 1, confirmed the predicted interference pattern. Responses were significantly slower for words with affectively incongruent phonetic forms (positive–frowning or negative–smiling, see Fig. 2) than for phonetically neutral words ($\beta = 39.4$ ms, $s.e. = 13.8$, $p = 0.004$). Responses for words with congruent phonetic forms (positive–smiling, negative–frowning) were equally fast as for neutral words ($\beta = 1.5$ ms, $s.e. = 13.8$, n.s.). RTs were faster for positive than for negative words ($\beta = 120$ ms, $s.e. = 33.5$, $p < 0.001$). Interactions were not significant, as confirmed by a likelihood ratio test [$\chi^2(2) < 1$, n.s.].

4. Discussion

In the first pretest, listeners' phonetic judgements of affective stimulus words were influenced by the semantic valence of those words, yielding more "smiling" ratings for semantically positive words and more "frowning" ratings for negative words, across phonetic manipulation conditions. This interference effect of semantic valence on phonetic ratings supports the main hypothesis in this study.

Conversely, in the main experiment, listeners showed a significant impairment in their comprehension of these spoken words, as measured by a semantic classification task, if the phonetic form of a word was incongruent with its semantic valence. Although some neurological and behavioral studies have suggested a general dissociation between emotional auditory processing and linguistic processing (e.g. Scott et al., 1997), our results indicate that vocal expressions of affect are integrated with linguistic properties of an utterance, thus adding affective redundancy. When redundancy is not provided (i.e., in incongruent conditions), the utterance is more difficult to understand.

Previous work in the literature (e.g. Kitayama and Ishii, 2002; Nygaard and Queen, 2008) investigated perceptual interference by varying affective congruency of prosody in natural speech tokens. Articulatory interference of acting speakers who simulated the intended emotions was a possible account for the effect. Namely, human actors may have more difficulty in producing natural tokens of affective words with incongruent emotions, as opposed to words with congruent emotions. To address this limitation, here synthetic speech was used. Therefore, the present interference effect cannot be attributed to articulatory interference in acted speech, nor can it be ascribed to visual affective expressions, because visual cues were absent, nor to priming effects (Klauer and Musch, 2003; Foroni and Semin, 2009) because effects were measured on the affected words themselves. Our results are most likely due to immediate resonance between the valence of the target word and the affective phonetic properties of that spoken word.

The present results are in line with embodied theories of language comprehension suggesting that listeners resonate to both acoustic and semantic information during spoken language comprehension (cf. Zwaan and Taylor, 2006). The affective resonance observed in the present experiment corresponds with the recently proposed Simulation of Smiles Model (Niedenthal et al., 2010; Hietanen et al., 1998; Dimberg et al., 2000; Niedenthal, 2007) which claims that the meaning of a smile is conveyed by means of motor mimicry of the observed smile – extended here to include not only visually but also phonetically observed smiles – somewhat analogous to the Motor Theory of Speech Perception (Galantucci et al., 2006). Similarly, imitation (i.e., speech mimicry) of a foreign accent improves comprehension of other utterances spoken in that accent (Adank et al., 2010), and listeners' facial expressions contribute to their recognition of auditorily presented emotions (Hawk et al., 2012). All these findings suggest that motor mimicry may contribute not only to speech comprehension, but also to affect perception. In our view, a perceived smile may elicit a smiling gesture (albeit a weak one) in the listener, which in turn interferes with speech comprehension through motor resonance (Zwaan and Taylor, 2006) and/or through affective resonance (Gallese, 2003, 2009).

5. Conclusion

In sum, comprehension of spoken language is not merely based on *what* is said, but also on *how* it is said – namely the affective facial expression coinciding with speech production. Speech comprehension is therefore an integrated process that benefits from the affective expressions that modulate how we say what we say. If speakers smile or frown while talking, then the audible effects of these affective cues influence our comprehension of spoken words.

Acknowledgements

We thank Sieb Nooteboom, Frank Wijnen, Marcel Schmeets and Karin Wagenaar for helpful comments and suggestions, and Theo Veenker and Iris Mulders for technical and logistical assistance.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.specom.2012.03.004.

References

Adank, P., Hagoort, P., Bekkering, H., 2010. Imitation improves language comprehension. Psychol. Sci. 21 (12), 1903–1909.

Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59 (4), 390–412.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E., 2003. How to find trouble in communication. Speech Commun. 40 (1–2), 117–143.

Boersma, P., Weenink, D., 2011. Praat: doing phonetics by computer [Computer software]. Retrieved from <http://www.praat.org>.

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., Maneewongvatana, S., 2008. Encoding emotions in speech with the size code. Phonetica 65 (4), 210–230.

Dimberg, U., Thunberg, M., Elmehed, K., 2000. Unconscious facial reactions to emotional facial expressions. Psychol. Sci. 11 (1), 86–89.

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

Foroni, F., Semin, G.R., 2009. Language that puts you in touch with your bodily feelings: the multimodal responsiveness of affective expressions. Psychol. Sci. 20 (8), 974–980.

Frick, R.W., 1985. Communicating emotion: the role of prosodic features. Psychol. Bull. 97 (3), 412–429.

Galantucci, B., Fowler, C.A., Turvey, M.T., 2006. The motor theory of speech perception reviewed. Psychon. Bull. Rev. 13 (3), 361–377.

Gallese, V., 2003. The roots of empathy: the shared manifold hypothesis and the neural basis of intersubjectivity. Psychopathology 36 (4), 171–180.

Gallese, V., 2009. Motor abstraction: a neuroscientific account of how action goals and intentions are mapped and understood. Psychol. Res. 73 (4), 486–498.

Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G., 1996. Action recognition in the premotor cortex. Brain 119, 593–609.

Grimshaw, G.M., 1998. Integration and interference in the cerebral hemispheres: relations with hemispheric specialization. Brain Cogn. 36 (2), 108–127.

Hawk, S.T., Fischer, A.H., Van Kleef, G.A., 2012. Face the noise: embodied responses to nonverbal vocalizations of discrete emotions. J. Pers. Soc. Psychol. 102 (4), 796–814.

Hietanen, J.K., Surakka, V., Linnankoski, I., 1998. Facial electromyographic responses to vocal affect expressions. Psychophysiology 35 (5), 530–536.

Kitayama, S., Ishii, K., 2002. Word and voice. Spontaneous attention to emotional utterances in two languages. Cogn. Emot. 16 (1), 29–59.

Klatt, D.H., 1980. Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am. 67 (3), 971–995.

Klauer, K.C., Musch, J., 2003. Affective priming: findings and theories. In: Musch, J., Klauer, K.C. (Eds.), The Psychology of Evaluation: Affective Processes in Cognition and Emotion. Lawrence Erlbaum Ass, Mahwah, NJ, pp. 7–49.

Kohler, E., Keysers, C., Umilta, M.A., Fogassi, L., Gallese, V., Rizzolatti, G., 2002. Hearing sounds, understanding actions: action representation in mirror neurons. Science 297 (5582), 846–848.

Lasarcyk, E., Trouvain, J., 2008. Spread lips + raised larynx + higher F0 = smiled speech? An articulatory synthesis approach. In: The 8th International Speech Production Seminar (ISSP), Strasbourg.

Mehrabian, A., Wiener, M., 1967. Decoding of inconsistent communications. J. Pers. Soc. Psychol. 6 (1), 109–114.

Neumann, R., Strack, F., 2000. “Mood contagion”: the automatic transfer of mood between persons. J. Pers. Soc. Psychol. 79 (2), 211–223.

Niedenthal, P.M., 2007. Embodying emotion. Science 316 (5827), 1002–1005.

Niedenthal, P.M., Mermillod, M., Maringer, M., Hess, U., 2010. The simulation of smiles (SIMS) model: embodied simulation and the meaning of facial expression. Behav. Brain Sci. 33 (6), 417–433.

Nygaard, L.C., Queen, J.S., 2008. Communicating emotion: linking affective prosody and word meaning. J. Exp. Psychol. Hum. Percept. Perform. 34 (4), 1017–1030.

Ohala, J.J., 1980. The acoustic origin of the smile. J. Acoust. Soc. Amer. S33. Retrieved from <http://linguistics.berkeley.edu/phonlab/users/ohala/papers/smile.pdf>.

Ohala, J.J., 1983. Cross-language use of pitch: an ethological view. Phonetica 40 (1), 1–18.

Paulmann, S., Titone, D., Pell, M.D., 2012. How emotional prosody guides your way: evidence from eye movements. Speech Commun. 54 (1), 92–107.

Pell, M.D., Paulmann, S., Dara, C., Alasseri, A., Kotz, S.A., 2009. Factors in the recognition of vocally expressed emotions: a comparison of four languages. J. Phonetics 37 (4), 417–435.

Quené, H., Van den Bergh, H., 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. J. Mem. Lang. 59 (4), 413–425.

Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech Commun. 40 (1–2), 227–256.

Schirmer, A., Kotz, S.A., 2003. ERP evidence for a sex-specific Stroop effect in emotional speech. J. Cognitive Neurosci. 15 (8), 1135–1148.

Schirmer, A., Kotz, S.A., Friederici, A.D., 2002. Sex differentiates the role of emotional prosody during word processing. Cognitive Brain Res. 14, 228–233.

Schröder, M., 2006. Expressing degree of activation in synthetic speech. IEEE Trans. Audio Speech Process. 14 (4), 1128–1136.

Scott, S.K., Young, A.W., Calder, A.J., Hellawell, D.J., Aggleton, J.P., Johnsons, M., 1997. Impaired auditory recognition of fear and anger following bilateral amygdala lesions. Nature 385 (6613), 254–257.

Stroop, J.R., 1935. Studies of interference in serial verbal reaction. J. Exp. Psychol. 18, 643–662.

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

Tesink, C.M.J.Y., Petersson, K.M., van Berkum, J.J.A., van den Brink, D., Buitelaar, J.K., Hagoort, P., 2008. Unification of speaker and meaning in language comprehension: an fMRI study. J. Cognitive Neurosci. 21 (11), 2085–2099.

Van Berkum, J.J.A., van den Brink, D., Tesink, C.M.J.Y., Kos, M., Hagoort, P., 2007. The neural integration of speaker and message. J. Cognitive Neurosci. 20 (4), 580–591.
Williams, C.E., Stevens, K.N., 1972. Emotions and speech: some acoustical correlates. J. Acoust. Soc. Am. 52 (4B), 1238–1250.
Wilson, S.M.W., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nat. Neurosci. 7, 701–702.
Xu, Y., Kelly, A., 2010. Perception of anger and happiness from resynthesized speech with size-related manipulations. In: The Speech Prosody 2010 Conference, Chicago. Retrieved from http://www.isca-speech.org/archive/sp2010/papers/sp10_027.pdf.
Zwaan, R.A., Taylor, L.J., 2006. Seeing, acting, understanding: motor resonance in language comprehension. J. Exp. Psychol. Gen. 135 (1), 1–11.