Journal of Memory and Language 88 (2016) 70–86

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Prosody conveys speaker's intentions: Acoustic cues for speech act perception

Nele Hellbernd *, Daniela Sammler

Otto Hahn Group "Neural Bases of Intonation in Speech", Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

ARTICLE INFO

Article history:
Received 7 May 2015
revision received 5 January 2016

Keywords:
Prosody
Intention
Speech acts
Acoustics
Pragmatics

ABSTRACT

Action-theoretic views of language posit that the recognition of others' intentions is key to successful interpersonal communication. Yet, speakers do not always code their intentions literally, raising the question of which mechanisms enable interlocutors to exchange communicative intents. The present study investigated whether and how prosody—the vocal tone—contributes to the identification of "unspoken" intentions. Single (non-)words were spoken with six intonations representing different speech acts—as carriers of communicative intentions. This corpus was acoustically analyzed (Experiment 1), and behaviorally evaluated in two experiments (Experiments 2 and 3). The combined results show characteristic prosodic feature configurations for different intentions that were reliably recognized by listeners. Interestingly, identification of intentions was not contingent on context (single words), lexical information (non-words), and recognition of the speaker's emotion (valence and arousal). Overall, the data demonstrate that speakers' intentions are represented in the prosodic signal which can, thus, determine the success of interpersonal communication.

Introduction

During conversations, humans regularly decode not only *what* is said but also *why* (Bühler, 1934; Grice, 1957; Wittgenstein, 1953). Depending on the latter, we may understand the same statement "It's hard to be punctual in the morning" as empathic concern, criticism, or simply as a matter of facts. Pragmatic theory posits that it is particularly the *why*—the communicative intention of the speaker—that drives the recipient's behavior and is the motive of communication. Yet, how intentions are (de)coded in interpersonal communication is still not fully understood. Contemporary pragma-linguistic theories posit that listeners identify the speaker's goal via pragmatic inference (Wilson & Sperber, 2012), taking conversation context and "common ground" (Clark & Carlson, 1981; Levinson, 2013; Stalnaker, 2002; Tomasello, 2005; Wichmann, 2002) into account. Alternatively, other studies seek to identify extralinguistic cues that reveal a speaker's intention, such as facial expressions (Fridlund, 1994; Frith, 2009; Parkinson, 2005), properties of biological motion (Di Cesare, Di Dio, Marchi, & Rizzolatti, 2015), or gestures (Bucciarelli, Colle, & Bara, 2003; Enrici, Adenzato, Cappa, Bara, & Tettamanti, 2011). The present study will focus on speech prosody—the tone of the voice—and will weigh its potential to convey communicative intentions.

The question of how interlocutors decode the *why* of an utterance is grounded in *action-theories of language*. In the middle of the 20th century, scholars like Bühler (1934), Wittgenstein (1953), or Grice (1975) recognized that language is more than strings of symbols that are understood

* Corresponding author at: Otto Hahn Group "Neural Bases of Intonation in Speech", Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1a, 04103 Leipzig, Germany. Fax: +49 341 9940 2204.

E-mail addresses: hellbernd@cbs.mpg.de (N. Hellbernd), sammler@cbs.mpg.de (D. Sammler).

http://dx.doi.org/10.1016/j.jml.2016.01.001
0749-596X/© 2016 Elsevier Inc. All rights reserved.

by retrieving their conventional, *coded* meaning. In their view, language is an *intentional action* and gains meaning through its employment. Utterances become instruments to influence the behavior of the interlocutor. The meaning of an utterance must be found in its underlying intention. It was Grice (1957) who particularly promoted the central role of intentions in communication. He advocated the idea that intentions drive speakers' behaviors (e.g., utterances) whose sole function is to have an effect on the addressee by virtue of having their intention recognized (cf. Levinson, 2006). Notably, the intention of the speaker—the *speaker meaning* in Grice's terms—not necessarily surfaces in the overt lexical content of the utterance, as shown in the example on punctuality above, but needs to be interpreted by the listener.

This idea later became central to speech act theory by Austin (1962) and Searle (1969) who considered utterances as actions—or *speech acts*—with specific interpersonal goals such as promising, apologizing, or warning. Like Grice, they claimed that speakers convey information on at least two levels: (1) the *propositional content* carrying the lexical meaning of *what* is said, and (2) the *illocutionary force* representing the action and speaker's intention—the *why*. As mentioned above, it is this second level—what the speaker is attempting to accomplish with a remark—that is thought to predominantly drive the interlocutor's (conversational) reaction. Notably, illocutionary force is often expressed implicitly (i.e. without the performative verb) or even indirectly, hence requiring some sort of inference on the part of the listener (Austin, 1962; Bach, 1994).

Interestingly, the notion of implicitness and indirectness conflicts with Grice's *cooperative principle* (1975), which describes principles for effective communication in conversation in four maxims. Following his *maxim of manner*, speakers ought to shape their utterances in ways that support the purpose of the conversation. Hence, speakers should produce unambiguous cues that make their intentions comprehensible to listeners. The fact that this seems often not to be the case but listeners still efficiently recognize the speaker's intent has fueled research on the cognitive and neural bases of comprehending communicative intentions. A great deal of work has focused on implicit speech acts, i.e. utterances that express the speaker's intention and illocutionary force without inclusion of the performative verb (e.g. "I will be there." expressing a promise without including the verb "promise"). These studies demonstrated the psychological reality of speech acts (Holtgraves, 2005), their automatic (Holtgraves, 2008a; Liu, 2011) and early recognition during conversation turns (Egorova, Pulvermüller, & Shtyrov, 2014; Egorova, Shtyrov, & Pulvermüller, 2013; Gisladottir, Chwilla, & Levinson, 2015), and their importance for conversation memory (Holtgraves, 2008b). However, despite their importance for understanding human communication, these studies remain incomplete in one particular way: They often rely on written linguistic material and, thus, miss out on extralinguistic cues that are usually available during natural spoken conversations. These cues comprise signals expressed via additional communicative channels like eyes, face, body, or voice and may render the speaker's intention less implicit and indirect than typically thought. The present study will focus on vocal acoustic cues, i.e., prosody, as one non-verbal channel in interpersonal conversation that may play an important role for speakers and listeners to express and recognize communicative intentions.

The term prosody refers to variations in pitch, loudness, timing, or voice quality over the course of an utterance (Warren, 1999) that can modify the communicative content of a message, both linguistically and paralinguistically (Bolinger, 1986). Linguistically, prosody has direct effects on the information structure of an utterance. It conveys, for example, semantic relationships (Cutler, Dahan, & van Donselaar, 1997; Wagner & Watson, 2010), disambiguates the syntactic constituent structure (Carlson, Frazier, & Clifton, 2009), and marks declarative vs. interrogative sentence mode (Sammler, Grosbras, Anwander, Bestelmeyer, & Belin, 2015; Schneider, Lintfert, Dogil, & Möbius, 2006; Srinivasan & Massaro, 2003). Paralinguistically, the "manner of saying" conveys additional information that goes beyond the linguistic content. Whether or not this includes intentions is a matter of debate (Bolinger, 1986) and will be topic of the present research.

Until now, most studies on paralinguistic prosody either focused on the speaker's emotion (Banse & Scherer, 1996; Bänziger & Scherer, 2005; Frick, 1985; Simon-Thomas, Keltner, Sauter, Sinicropi-Yao, & Abramson, 2009) or, more recently, on their attitude, for example, the politeness, confidence, or sincerity of the speaker (Jiang & Pell, 2015; Monetta, Cheang, & Pell, 2008; Rigoulot, Fish, & Pell, 2014) and often sought to determine links between the acoustics of the prosodic signal and the listeners' comprehension of the paralinguistic message. Although opinions diverge on whether prosody as such can convey meaning, i.e. without contextual information (see below) (Cutler, 1976; Wichmann, 2000, 2002), studies revealed distinct acoustic properties for the prosodic expression of different emotions (Banse & Scherer, 1996; Szameitat, Alter, Szameitat, Darwin, et al., 2009; Szameitat, Alter, Szameitat, Wildgruber, et al., 2009) and attitudes (Blanc & Dominey, 2003; Morlec, Bailly, & Aubergé, 2001; Uldall, 1960). Similarly, on the perception side, researchers showed that participants were able to identify the speaker's attitude (Morlec et al., 2001; Uldall, 1960) and emotion by prosodic differences alone, in verbal (Banse & Scherer, 1996; Morlec et al., 2001) and non-verbal utterances (Monetta et al., 2008; Sauter, Eisner, Calder, & Scott, 2010), in laughter (Szameitat, Alter, Szameitat, Darwin, et al., 2009; Szameitat, Alter, Szameitat, Wildgruber, et al., 2009), and to some extent even cross-culturally (Sauter, Eisner, Ekman, & Scott, 2010).

Compared to this active field of research, only little is known about the perceptual reality, relevance and effectiveness of prosodic cues in conveying *intentions*. We consider communicative intentions as the goals of interpersonal actions (e.g., language) that are meant to be recognized by the interlocutor and to influence her (conversational) reactions. This differentiates communicative intentions from basic emotions that do not necessarily need another person to be displayed, and attitudes that are not necessarily meant to purposefully influence conversation partners (Wichmann, 2000). Certainly, both emotions and attitudes can be expressed for communicative purposes (Fridlund, 1994; Mead, 1934; Parkinson, 2005) and often take an effect on the listener by virtue of their "expressive function" (Bühler, 1934). Yet, their intended goal remains rather underspecified compared to the "specific intentions for specific turns" (cf. Holtgraves, 2008a) proposed by action-theoretic accounts of language, particularly by speech act theory (Austin, 1962; Searle, 1969).

To date, the role of prosody for the non-literal expression and recognition of different intentions still lacks detailed investigation, although several findings from developmental studies and psycholinguistics point to the relevance of extralinguistic vocal cues in intentional communication. For quite some time, studies on intonational development have been focusing on the emergence of illocutionary skills in infants, considering intonation patterns as primitive devices that preverbal infants use to express their communicative intentions (Dore, 1975). For example, 7- to 11-month-old babies were found to vocally distinguish between communicative and investigative (Papaeliou, Minadakis, & Cavouras, 2002) or emotional functions when babbling (Papaeliou & Trevarthen, 2006). This competence was proposed to regulate cooperative interactions with their parents as a prerequisite for language acquisition. Furthermore, infants' intonations of babble at the end of their first year (Esteve-Gibert & Prieto, 2013), or words in their second year of life (e.g., Furrow, Podrouzek, & Moore, 1990; Marcos, 1987; Prieto, Estrella, Thorson, & Vanrell, 2012) were found to differ between simple speech acts such as complaining, requesting, or greeting. These combined findings were taken as evidence for a prosodic choice that prelinguistic infants make to communicate their intentions (illocutions) while their propositional (locutionary) abilities are still limited. One challenge that these studies have to face, though, is their dependency on adult, post hoc interpretations of infants' vocal actions that are usually based on the context in which the vocalizations were produced. This bears the risk that raters—although experts (e.g., mothers or phoneticians)—might overestimate or misinterpret the children's (true) motives or draw conclusions from cues other than prosody. Studies with adult speakers who can report on their intentions are necessary to corroborate the link between prosody and communicative intentions, and to show its persistence in adulthood.

The present study aimed to fill this gap by conceptualizing a speaker's intention in terms of speech acts (Austin, 1962; Holtgraves, 2002; Searle, 1969) and investigating the role of prosody in decoding illocutionary force. Note that it was not our goal to describe the prosody of a complete set of speech acts or to investigate the reality of speech act theory. Rather, we aimed to demonstrate that—in identical utterances pronounced according to a limited set of intentions—speakers produce well-identifiable characteristic prosodic patterns, and that these patterns can be reliably recognized by listeners. This adds to the debate whether prosody can convey meaning on its own, i.e., may be conventionalized for different communicative concepts. Alternative views regard prosody as a contrastive marker that does not carry meaning by itself but signals the presence of "unspoken" meaning by deviating from normal prosody, and hence, motivates listeners to infer the implied message by taking context information into account (Cutler & Isard, 1980; Levinson, 2013). Here, we tested the hypothesis that prosodic patterns as such can be sufficiently distinct, to a degree that listeners can recognize the broad communicative concept and intention in the prosodic speech signal. Therefore, our stimulus set comprised single words and non-words, i.e., tokens free of context and lexical meaning, that were pronounced with six different intonations representing the speech acts criticism, doubt, naming, suggestion, warning, and wish. In three experiments, we combined acoustic analyses of these speech signals with perceptual judgments of listeners (for a similar approach, see Banse & Scherer, 1996; Sauter, Eisner, Calder, et al., 2010). If prosody itself codes speakers' intentions, different speakers should employ similar cue configurations when conveying the same intention, and participants should be able to recognize the intention without contextual information (i.e., in single words) and irrespective of whether the speech sound carries lexical meaning or not (i.e., in words and non-words).

One important consideration for our investigations of communicative intentions in prosody is the relation to emotional components in the speaker's tone of voice. Although we advocated a conceptual differentiation of intentions and emotions above, we have to keep in mind that emotions (e.g., fear) might drive intentions (e.g., to warn the interlocutor). Hence, both may be intertwined in the production and perception of communicative utterances. In an attempt to show that the comprehension of intentions is more than the recognition of emotions or affect in the prosodic signal, we further assessed the valence and arousal of our speech stimuli according to dimensional models of affect (Remmington, Fabrigar, & Visser, 2000; Wundt, 1896) (we will use the term emotion throughout the text to refer to these affect measures). These values were then used to correct the perceptual recognition of intentions for the contribution of emotion (see below).

The present study took three steps: We started with analyses of the acoustics of speakers' vocal expressions of speech acts by means of discriminant analyses (Experiment 1). If speech acts as carriers of intentions are coded in characteristic prosodies (i.e. show some consistency of the prosodic pattern across speakers and across tokens within speakers), it should be possible to classify the different categories of speech acts based on their acoustic features alone, in words and non-words alike. Second, we tested whether listeners are able to identify the correct intention based on the prosodic pattern alone (Experiment 2) in a 6-alternative forced choice (6-AFC) categorization task and ratings of the stimuli on every speech act scale (e.g., "How much does it sound like criticism?"). If prosody conveys meaning in a partly conventionalized way, listeners should be able to classify the intentions despite lack of context (i.e., in single words) and irrespective of lexical meaning (i.e., similarly in words and non-words). Finally, we determined which acoustic parameters contribute most to the perception of the respective intention (Experiment 3). Therefore, we fed the acoustic parameters into multiple regression analyses to predict the participants' ratings on each speech act scale. Furthermore, to control for a possible influence of emotion on intention recognition, the regression analyses were repeated once after valence and arousal ratings of the stimuli had been regressed out.

In summary, the present study sought to demonstrate that prosody carries information about the speaker's communicative intention by (i) identifying characteristic prosodic feature configurations of a set of speech acts that are (ii) reliably recognized by listeners, (iii) despite the lack of context information (single words) and semantic content (non-words) and the control for emotional processing of the stimuli.

Experiment 1 – acoustics

The goal of Experiment 1 was to investigate whether speakers use characteristic acoustic features to convey their intentions. If so, it should be possible to classify the speech stimuli into the corresponding speech act categories based on their acoustic features alone and irrespective of word meaning. Specifically, we focused on duration, intensity, pitch, and spectral features that have been analyzed in similar approaches in emotion research (e.g. Banse & Scherer, 1996; Blanc & Dominey, 2003; Sauter, Eisner, Calder, et al., 2010). In such studies, pitch cues were predominant when emotions were expressed verbally, compared to a stronger weighting of spectral features in non-verbal utterances, making it likely that pitch cues will play a major role in the present experiment.

For the current study, four speakers produced single-word stimuli with varying prosodies to express six different intentions, i.e., the speech acts criticism, doubt, naming, suggestion, warning, and wish. To obtain stimuli that are representative for typical language use, all speakers were non-actors, i.e., they relied on their intuition—not training in acting—to express the intention in a way that could be understood by an imaginary interlocutor. For high stimulus quality, all speakers were, however, familiar with sound recordings, i.e., working as voice coaches or speech scientists. This choice of professional speakers with only minimal training in acting is an attempt to face the criticism, first raised in emotion research, that actors' prosodic patterns may deviate from those used in everyday conversations (Jürgens, Hammerschmidt, & Fischer, 2011; see also General discussion). Apart from that, it should be mentioned that intentions are typically expressed more voluntarily than emotions and are, hence, less dependent on the spontaneity of the utterance. Altogether, the present stimuli were recorded such to grant generalizability of the results to natural language use.

Materials and methods

Ethics approval

The ethics committee of the University of Leipzig, Germany approved the present and all following experiments in this study.

Stimulus recordings

Four trained native German speakers (voice coaches, 2 female) were invited to record the German words "Bier" (*beer*) and "Bar" (*bar*) as well as the non-words "Diem" and "Dahm" (for examples, see Appendix D: Supplementary material). These (non-)words were intoned to express six different communicative intentions or speech acts: criticism, doubt, naming, suggestion, warning, and wish. The chosen speech acts were plausible for our stimulus words "beer" and "bar" and fit into the broader speech act categories as defined by Searle and Vanderveken (1985). To elicit the respective intentions in the speakers, they read short scenarios that described a situation in which they interacted with an interlocutor (see Appendix A). They were allowed to utter an initial sentence and to vocalize freely until they felt ready to articulate the intention shortened to the single essential word. This recording approach, instead of using natural speech recordings, was chosen to obtain clear portrayals of intentions in good sound quality. Recordings were conducted in a soundproof room with the microphone (Rode NT55) approximately 20 cm in front of the speaker and digitized at a 44.1 kHz sampling rate in a 16-bit mono format. The words and non-words were repeated several times to obtain eight variants per stimulus in good quality. The resulting stimulus set, thus, comprised 768 stimuli, with eight repetitions of four (non-) words expressed as six speech acts by four speakers.

Acoustic features

For investigating acoustic features of the speech acts, we obtained seven acoustic measures that are commonly used in experiments on human voice and speech stimuli (e.g., Banse & Scherer, 1996; Sauter, Eisner, Calder, et al., 2010). Using Praat software (Boersma & Weenink, 2014) we extracted the number of voiced frames as a measure of stimulus duration, mean intensity, harmonics-to-noise ratio (HNR), mean fundamental frequency (f0) as well as pitch rise, measured as the difference between offset and onset f0. Furthermore, we extracted the spectral center of gravity and the standard deviation of the spectrum. The mean acoustic characteristics as measured with Praat are presented in Table C1 (Appendix C). Statistical analyses showed that speakers had used very similar acoustic cues to intone speech acts in words and non-words. *T*-tests for paired samples comparing the acoustics of words and non-words for each speech act category were largely non-significant. Only exception were HNR and spectral center of gravity that showed differences in some, but not all speech act categories (see Table C2 in Appendix C for more details). These differences are, however, likely to be caused by the different consonants ("r" in words vs. "m" in non-words) rather than by differences in prosody.

Discriminant analyses

Discriminant analyses were performed for words and non-words separately, with the seven acoustic features as independent variables and the speech act category (criticism, doubt, etc.) as dependent variable. These analyses sought to identify linear functions of acoustic feature combinations that maximize differences between speech act categories. In other words, these analyses tested whether

Table 1
Results of cross-validated (jackknife) discriminant analysis for classification of speech acts from the acoustic features (in %). Correct classifications are shown in bold.

Stimulus type	Classification					
	Criticism	Doubt	Naming	Suggestion	Warning	Wish
Words						
Criticism	**76.6**	4.7	0	10.9	7.8	0
Doubt	1.6	**92.2**	0	1.6	0	4.7
Naming	0	0	**100**	0	0	0
Suggestion	3.1	1.6	0	**95.3**	0	0
Warning	0	0	0	0	**100**	0
Wish	0	0	9.4	0	0	**90.6**
Non-words						
Criticism	**79.7**	6.2	0	7.8	4.7	1.6
Doubt	7.8	**84.4**	3.1	0	0	4.7
Naming	0	0	**100**	0	0	0
Suggestion	1.6	0	0	**98.4**	0	0
Warning	0	0	0	0	**100**	0
Wish	0	0	4.7	0	0	**95.3**

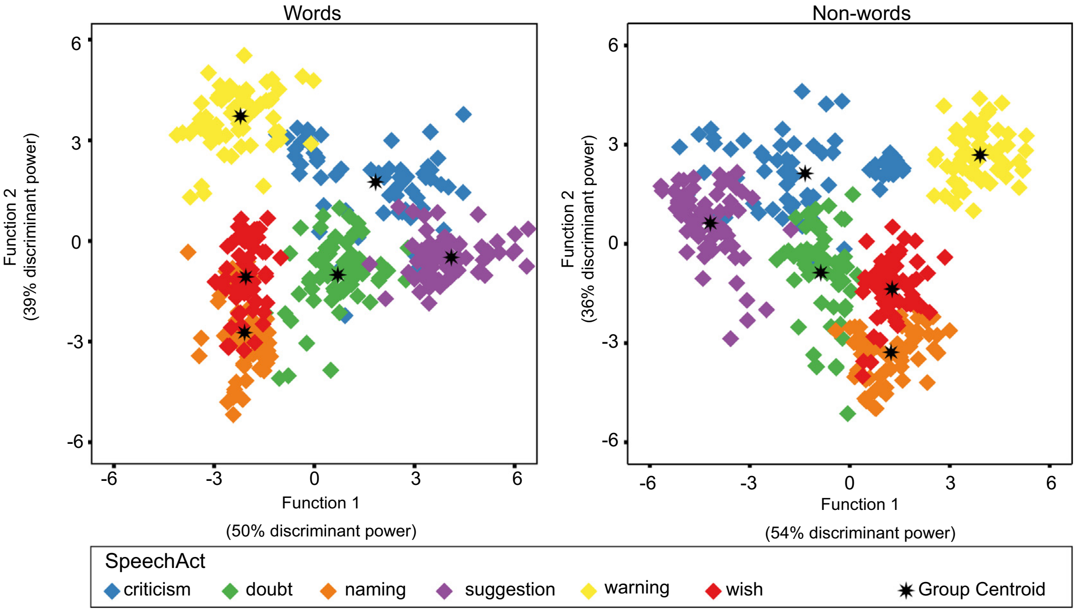

Fig. 1. Results of the discriminant analyses plotted along the first and second discriminant function. Data points correspond to stimuli of the six speech act categories.

the acoustic features alone have sufficient discriminant power to reliably group the stimuli that express the same intention. The discriminant analyses for both words and non-words were cross-validated with a jack-knife procedure, and the distribution of the results was validated with chi-square tests.

Results

The discriminant analyses classified the correct speech act category for 92% of all word stimuli and 93% of all non-word stimuli. These results are highly above chance-level (17%) as was tested with chi-square tests: $\chi^2(35) = 1717.6$, $p < .001$ for words and $\chi^2(35) = 1702.6$, $p < .001$ for non-words. Classification results of the discriminant analyses for the different types of speech acts are demonstrated as confusion matrices in Table 1. The highest results were found for naming and warning (both 100% correct classification for words and non-words), while the lowest results were obtained for criticism in words (76.6% correct classification), and non-words (79.7%). Additional chi-square tests showed that the discriminant model classified our stimuli better than chance (chance-level: 17%) for every type of speech act ($\chi^2(5) > 151$, p's < .001). Fig. 1 shows the classification of the different speech acts by the first two discriminant functions. The first function (x-axes) explained 49.6% of the variance for words and 54.2% for non-words and was mainly based on the acoustic measure of pitch rise (offset–onset f0). The second discriminant function (y-axes) had an additional discriminant power of 38.6% for words and 36.4% for non-words and was most related to the mean intensity and mean f0 of the stimulus (see Table C3, Appendix C). Additionally, a third function (not depicted in Fig. 1 for reasons of clarity) explained 10.5% of variance for words and 7.7% for non-words and showed highest correlation with the duration of the stimuli. The last two discriminant functions from our analyses explained only minor effects (function 4: 0.9% for words and 1.2% for non-words, function 5: 0.4% for words and 0.5% for non-words) and were neglected from further investigations.

Discussion

The acoustic features of our stimulus set could be used to accurately classify the correct speech act, for words and non-words alike. This demonstrates the distinctiveness of the prosodic patterns that speakers deliberately applied to code their intentions in the tested speech act categories and the relative independence of prosody from lexical content. Furthermore, the high accuracy of the classification implies a reasonable consistency of the relevant prosodic cues across speakers and utterances, and may point to the existence of feature configurations that speakers consider conventional and appropriate for different communicative goals. For example, the warning stimuli were loudest and had the most arched pitch contour with a salient peak in the middle of the word as is appropriate for the urgent nature of a warning. In comparison, the naming stimuli showed the least salient acoustic features with low mean pitch, flat pitch contour, low intensity and little spectral variation in line with the neutral character of the expression. As expected, pitch rise and mean f0, together with mean intensity were the most influential acoustic features in these analyses, while spectral features had only weak discriminant power. In sum, the data show that speakers can use prosody as a channel of communication to convey their intentions. Note that we do not expect that speakers possess different prosodic patterns for all possible intentions or speech acts. Yet, we believe that speakers choose salient, distinguishable and probably culturally learned prosodic signatures to trigger cognitive processes in the addressee to infer the communicative intent of the speaker beyond the overt lexical meaning.

Experiment 2 – behavior

After finding consistent acoustic differences between prosodic speech act expressions, we were interested in participants' perception of the stimuli. We investigated whether participants would be able to identify the different intentions based on the prosodic information in a 6-alternative forced-choice (6-AFC) categorization task. Participants, further, judged the valence and arousal of every stimulus (Remmington et al., 2000; Russell, 1980; Wundt, 1896) in the second half of the experiment, which allowed us to assess in how far speech acts may be classified based on their emotional tone.

Materials and methods

Participants

Ten participants were presented with the word stimuli (6 female, mean age ± *SD*: 24.6 ± 4.9), ten other volunteers performed the task with the non-word stimuli (4 females, mean age ± *SD*: 24.9 ± 2.6). We tested separate groups of participants for words and non-words to avoid transfer of the semantic meaning to the non-word stimuli. All participants reported normal hearing ability, gave written informed consent and were paid 7€ per hour for their participation.

Design and procedure

In the first half of the experimental session, participants were asked to assign each stimulus to one of the six possible speech act categories (criticism, doubt, naming, suggestion, warning, or wish). After having read short definitions for the different speech acts (Appendix B), they heard each sound stimulus once via headphones and were instructed to press the keys 1–6 on a keyboard. The speech act labels with corresponding numbers were displayed on a computer screen throughout the experiment. No feedback for the correctness of the response and no time limits were given. The experiment was separated into four blocks—one for each speaker. Block order and stimulus order within each block were pseudo-randomized by preallocating the speech acts with balanced probabilities. Chi-square tests were performed to test for above-chance classification across all speech acts and within single speech act categories.

In the second half of the experimental session, participants were asked to evaluate the valence and arousal of each stimulus. Therefore, they listened to the same stimuli again, in the same order as before. After each sound, they saw two visual analogue scales on the screen, first for valence (positive/negative), then for arousal (calm/excited), and placed their ratings with a continuous slider. The scales showed the outermost pictures of the Self-Assessment Manikin (Bradley & Lang, 1994) at the margins. No time constraint was given for the answers. Friedman tests were calculated to examine differences in the affect ratings among the speech act categories.

Results

Speech act categorization

In the 6-AFC task, participants were able to identify the correct speech act category of our stimuli with high accuracy for words (mean ± *SD*: 82 ± 13%) and for non-words (73 ± 17%)—with no significant difference between the participant groups for words and non-words ($t(18) = 1.26$, $p = .22$). Chi-square tests showed that participants' classification of every speech act category was better than being predicted by chance (chance level: 17%), both for words ($\chi^2(5) > 1082$, p's < .001) and non-words ($\chi^2(5) > 798$, p's < .001).

Confusion matrices for words and non-words are presented in Table 2. As with the acoustic analyses, the identification of criticism was lowest among the six speech act categories. For words as well as non-words, participants misclassified criticism most often as being doubt, and to a lower extent as warning. Furthermore, common confusions of the non-word stimuli were found for suggestion taken as doubt, wish, or criticism. To some extent,

Table 2
Behavioral categorization of speech acts (in %). Correct categorizations are shown in bold.

Stimulus type	Participants' responses					
	Criticism	Doubt	Naming	Suggestion	Warning	Wish
Words						
Criticism	**62.0**	24.0	0.2	4.9	6.9	2.0
Doubt	5.3	**83.4**	3.1	4.4	0.6	3.3
Naming	1.7	0.3	**90.0**	1.3	0.8	5.9
Suggestion	4.5	9.2	3.3	**80.3**	0.5	2.2
Warning	4.1	0	0.3	0.5	**89.5**	5.6
Wish	1.9	0.8	8.3	4.4	1.1	**83.6**
Non-words						
Criticism	**52.4**	29.7	1.1	3.9	9.4	3.4
Doubt	2.7	**82.6**	2.2	8.5	0	4.1
Naming	17.7	1.9	**74.1**	2.3	0.3	3.8
Suggestion	9.7	12.8	3.3	**63.7**	0.5	10.0
Warning	3.8	0.5	0.0	0.2	**94.2**	1.4
Wish	5.3	1.6	8.9	14.2	0.6	**69.4**

participants also misclassified wish as suggestion, and naming as criticism.

Emotion ratings

For the perception of emotion, mean ratings for valence and arousal differed significantly between the speech act categories, for words (valence: $\chi^2(5) = 35$, arousal: $\chi^2(5) = 43$, p's < .001) and non-words (valence: $\chi^2(5) = 44$, arousal: $\chi^2(5) = 45$, p's < .001). The results were very similar for words and non-words in each speech act category (Fig. 2). On the valence scale, the speech acts warning and criticism were perceived most negatively, whereas wish and suggestion were associated with a more positive valence. Doubt and naming were rated neutrally with regard to valence. The perception of the speakers' arousal was very calm for naming, wish, and doubt, and very excited for warning and criticism. Suggestion stimuli were rated in the middle range for arousal.

Discussion

Participants were well able to identify the speaker's communicative intention from the prosody alone as indicated by the highly significant results in the 6-AFC categorization task. Importantly, participants were able to make use of the prosodic signal with minimal context descriptions and without lexical content (see below). These data show that prosody is a powerful communicative channel that is used by listeners to decode the "unspoken" meaning and intention of the speaker and that may determine their respective conversational reaction. Interestingly, criticism was identified least reliably and was specifically confused with doubt, in line with the similar acoustic features of these two speech acts (Fig. 1, Table 1). It is well conceivable that the acoustic similarity of criticism and doubt may amount from their conceptual similarity—a rather depreciative stance toward an inner or outer event—a fact that

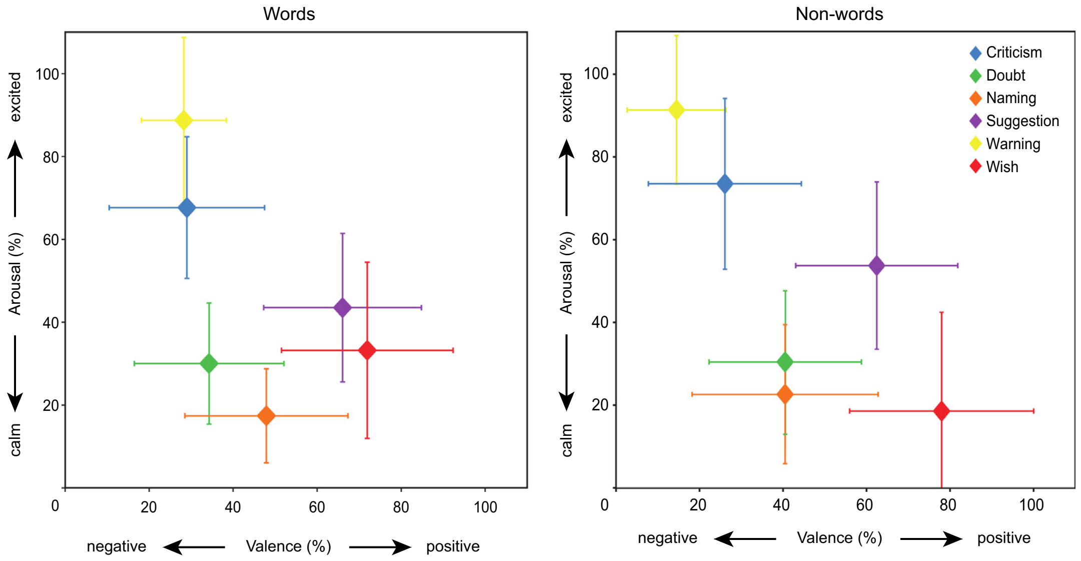

Fig. 2. Emotion ratings. Average scores of the valence and arousal ratings for each speech act category.

would further illustrate the intricate link between communicative intentions and prosody.

The valence and arousal ratings revealed distinctive affective properties of the different speech act categories, which were consistent across words and non-words (Fig. 2). This suggests that speaker's intentions may have emotional connotations that listeners are able to detect in the prosodic signal. A natural question that comes up is, how strongly the identification of intentions in prosody *depends* on emotion recognition and whether intentions can be recognized without taking emotion into account. We addressed this issue in Experiment 3.

Experiment 3 – behavior and acoustics

Experiment 1 and 2 demonstrated that prosodically coded intentions can be differentiated (i) physically based on characteristic acoustic feature configurations as well as (ii) perceptually in a 6-AFC task. What remains to be shown is, in how far the acoustic differences account for participants' ability to identify the speaker's intention. If participants use the prosody's acoustic information for intention understanding, it should be possible to predict listeners' perception from the acoustic measures and, further, to identify the different feature combinations that evoke specific speech act impressions. We addressed this question by feeding acoustic measures and typicality ratings for every speech act into a multiple regression analysis. Moreover, to assess the influence of emotion perception on intention recognition (see Experiment 2), we conducted an additional regression analysis in which valence and arousal ratings were regressed out.

Materials and methods

Participants

A new group of 20 healthy volunteers (10 females, mean age ± *SD*: 24.8 ± 4.1 years) for the words and 20 participants for the non-words (10 females, mean age ± *SD*: 24.6 ± 3.2 years) took part in a rating study. All participants reported normal hearing ability, gave written informed consent and were paid 7€ per hour for their participation.

Design and procedure

In this experiment, participants were asked to indicate to what extent each stimulus sounded like a given speech act category (criticism, doubt, naming, suggestion, warning, or wish). Compared to the 6-AFC categorization task, such speech act ratings provide a more refined and less strategy dependent measure for the participants' perception and allowed for the application of multiple regression analyses. In total, each stimulus was presented six times, once for every speech act scale, in separate blocks. Responses were given with a slider on a visual analogue scale from 0 to 100 ('intonation does not fit the intention at all' to 'intonation fits the intention very well'). Each block comprised the same 192 stimuli—four tokens of two (non-)words expressed as six speech acts by four speakers—that were chosen from the full stimulus pool of 384 stimuli. Stimuli and block order were again pseudo-randomized. The timing of the experiment was self-paced and participants were able to take breaks between blocks. The results of the ratings were analyzed by repeated-measures ANOVAs with Greenhouse–Geisser correction for the factor SPEECH ACT for every speech act scale separately.

Multiple regressions

To elucidate which acoustic features guided the participants' ratings on the speech act scales, we performed linear multiple regression analyses. Specifically, we used the acoustic features as predictors (independent variable) for the subjective ratings of the 192 stimuli (dependent variable), separately for words and non-words. Acoustic features were the same as in Experiment 1 and were chosen such to include measures of duration, intensity, pitch, and spectrum while keeping multicollinearity low (variance inflation factor words: <3.526, non-words: <4.561).

Furthermore, to demonstrate that intention perception is not merely determined by perceived emotional connotations in the speech signal, the emotion perception of the stimuli was regressed out in two steps: Firstly, separate regressions were calculated with the valence and arousal ratings as independent variables and the single speech act ratings as dependent variables. This way, we bound all the variance in the perceived speech act that could be explained by potentially perceived emotions. Thus, the residuals of these regressions should contain information about the participants' intention perception devoid of the perceived valence and arousal. Following this, new regressions were performed, now with the acoustic features as independent and the standardized residuals of the speech act ratings as dependent variables.

Results

Multiple regressions

The mean ratings of the stimuli according to the six different speech act scales are shown in Table 3. As can be seen in the diagonal, the highest ratings were obtained for the correct speech act category. This was confirmed by a significant main effect of SPEECH ACT in repeated-measures ANOVAs performed for every speech act scale separately (word stimuli: F's > 46.289, p's < .001; non-word stimuli: F's > 29.326, p's < .001). Post-hoc paired comparisons with Bonferroni correction showed that speech act stimuli were rated significantly higher on their corresponding scale than any other speech act category with p's < .03. Altogether, the ratings replicate the findings in the 6-AFC categorization task, in that also this new group of participants was well able to recognize and evaluate the speech acts correctly.

To examine whether specific patterns of acoustic features can predict subjective evaluation of the different speech act stimuli, the ratings together with the acoustic measures for the single stimuli were entered into multiple regression analyses, separately for each speech act rating scale. These regressions yielded highly significant results on all scales (see Table 4 for detailed results). The variance explained by the regression models ranged from 11.6% for the wish ratings to 52.7% for the warning ratings of the

Table 3
Participants' ratings of speech acts (min = 0, max = 100). Ratings on corresponding speech act scale are shown in bold.

Stimulus type	Speech act scale					
	Criticism	Doubt	Naming	Suggestion	Warning	Wish
Words						
Criticism	**70.0**	56.2	9.4	18.5	27.5	16.5
Doubt	45.7	**82.5**	13.7	16.7	10.7	9.5
Naming	12.2	13.2	**85.3**	10.5	9.3	17.4
Suggestion	15.7	28.0	17.7	**77.3**	7.7	26.9
Warning	22.8	18.6	8.2	16.0	**86.0**	29.9
Wish	6.5	10.1	21.3	16.3	6.0	**80.4**
Non-words						
Criticism	**63.3**	45.5	11.3	21.3	31.3	13.3
Doubt	23.0	**77.5**	17.2	22.7	6.1	20.6
Naming	14.9	13.1	**78.0**	17.7	8.7	21.7
Suggestion	20.5	35.3	24.9	**70.6**	10.9	18.3
Warning	26.0	5.4	7.5	8.5	**94.1**	9.1
Wish	10.1	11.4	25.7	22.8	5.4	**75.9**

Table 4
Multiple regression analyses of acoustic features and speech act ratings (beta-weights).

Acoustic parameter	Speech act ratings					
	Criticism	Doubt	Naming	Suggestion	Warning	Wish
Words						
Voiced frames	0.276***	0.348***	−0.420***	−0.227***	0.073***	0.083***
Mean f0	0.371***	0.402***	−0.326***	−0.211***	0.385***	−0.294***
Offset–onset f0	0.224***	0.315***	−0.277***	0.524***	−0.387***	−0.135***
Mean intensity	−0.015	−0.250***	−0.266***	0.368***	0.343***	0.118***
Mean HNR	−0.171***	−0.106***	0.141***	−0.008	−0.127***	0.044
Center of gravity	−0.083**	−0.106***	0.097***	−0.059**	−0.039*	0.042
SD spectrum	−0.085***	−0.082***	0.014	0.035*	−0.107***	0.149***
Adj R^2	**0.171*****	**0.273*****	**0.364*****	**0.349*****	**0.527*****	**0.116*****
Adj R^2 (emo-corr)	**0.153*****	**0.285*****	**0.228*****	**0.305*****	**0.162*****	**0.175*****
Non-words						
Voiced frames	0.183***	0.237***	−0.388***	−0.184***	0.056***	0.141***
Mean f0	0.121***	0.331***	−0.376***	−0.258***	0.289***	−0.204***
Offset–onset f0	0.259***	0.416***	−0.206***	0.407***	−0.387***	−0.062**
Mean intensity	0.168***	−0.315***	−0.291***	0.250***	0.479***	−0.100***
Mean HNR	−0.052*	0.033***	0.249***	−0.093***	−0.135***	−0.056*
Center of gravity	0.052*	−0.074***	0.199***	−0.182***	0.015	−0.077**
SD spectrum	0.105***	0.034***	−0.033***	−0.027***	−0.145***	0.175***
Adj R^2	**0.135*****	**0.269*****	**0.302*****	**0.260*****	**0.621*****	**0.176*****
Adj R^2 (emo-corr)	**0.092*****	**0.305*****	**0.132*****	**0.159*****	**0.270*****	**0.142*****

Beta weights and adjusted R^2 are depicted for multiple regressions using acoustic features as predictors and speech act ratings as dependent variables. Additionally, adjusted R^2 values are depicted after controlling for emotion perception (emo-corr). f0 = fundamental frequency; HNR = harmonics-to-noise ratio; *SD* = standard deviation; Adj = adjusted; emo-corr = overall performance of the multiple regressions after affective ratings had been regressed out.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

words and from 13.5% for the criticism ratings to 62.1% for warning ratings of the non-words. The beta weights of the regression functions indicate the degree to which the acoustic parameters predicted the ratings. Put differently, high absolute values of the beta weights reflect the importance of the corresponding acoustic feature for the prediction of the regression model. Almost all acoustic features contributed significantly to the predictions of the speech act ratings (see Table 4). While the spectral features (center of gravity and standard deviation of the spectrum) as well as the HNR yielded very low beta values in general (all <0.2, except for HNR in naming ratings in the non-words), the acoustic measures of pitch, amplitude, and duration reached absolute beta values of up to 0.524, suggesting that these parameters are key features for the comprehension of the intentions.

Fig. 3 shows the beta weights of the main acoustic features and reveals specific patterns of acoustic parameters for the prediction of the different speech act ratings: While high ratings for criticism and doubt were mainly predicted for long stimuli with high mean pitch and a rising pitch contour (positive beta weights for voiced frames, mean f0 and offset–onset f0), high ratings for naming were associated with short and soft stimuli with low mean pitch and falling pitch contour (negative beta weights for these measures). Suggestion ratings relied on short stimuli with low

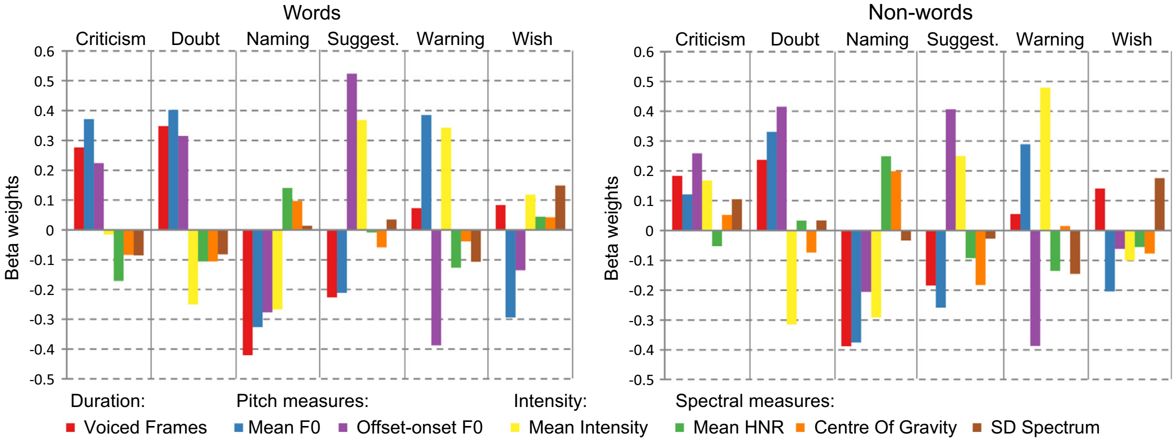

Fig. 3. Results of the multiple regressions for each speech act scale (columns). The bars represent the beta-weights for the seven acoustic features indicating how strongly they predicted the speech act rating.

mean pitch, but a strong pitch rise and high intensity. High ratings for warning were predicted, if the stimuli had a high mean pitch and intensity with a negative f0 offset–onset relation. Beta values in predictions for wish ratings showed the least clear pattern of acoustic information. For these stimuli, a low mean pitch was the most informative parameter. In total, the judgment of different speech act categories seems to be based on distinct acoustic patterns. Crucially, predictions of criticism and doubt ratings were based on similar patterns that only differed in the contribution of intensity. This is in line with the confusion of criticism and doubt in the behavioral categorization and ratings (Tables 2 and 3). Overall, the pitch-related features (mean f0 and offset–onset f0) had a high influence on the ratings in all speech act categories, qualifying them as the most important acoustic features in these analyses. A subset of speech acts was further influenced by amplitude and duration features, while spectral parameters only seemed to play a minor role. Notably, the results of the multiple regressions were very similar for the words and non-words, even though the analyzed data were not only based on a different stimulus set, but also on the ratings of two independent groups of participants. Therefore, these results validate the distinct acoustic patterns that shape the perception of different speech acts and intentions in single-word utterances.

Multiple regressions controlled for emotion

The second regression approach was performed to control for all variance that could be explained by the perception of valence and arousal. After an initial regression to predict speech act ratings as dependent variables from emotion ratings as independent variables, we conducted a second regression to explain residual speech act information by the acoustic measures. These regressions were still highly significant on all speech act scales (p's < .001, Table 4) and explained variance in the range from 15.3% for criticism ratings to 30.5% for suggestion ratings of the words, and from 9.2% for criticism ratings to 30.5% for doubt ratings of the non-words. Compared to the original regressions, there was a noticeable decrease in explained variance for naming and warning in words and non-words, which indicates that some of the variance could be explained by the emotion perception of these speech acts. Prediction of the other speech acts was virtually unchanged.

Discussion

The current experiment confirmed a link between the acoustics and perception of the speech act stimuli by multiple regression analyses in which distinct acoustic feature configurations significantly (but not fully) predicted the listeners' perception of the speech acts. The amount of variance explained (ranging between 12% and 62% depending on speech act type) was overall comparable to estimates found in previous studies on emotional prosody (Banse & Scherer, 1996; Sauter, Eisner, Calder, et al., 2010), validating our approach. In general, pitch features (mean pitch and offset–onset f0) were most influential for the perception of different intentions. Further important cues could be derived from intensity and duration measures, whereas spectral features contributed least to the intention predictions.

The amount of variance explained by the regressions was significant, but the values for some speech acts (e.g. 12% for wish in words or 13% for criticism in non-words) suggest that the acoustic features chosen for the analyses are not the only basis for intention perception. The inclusion of additional acoustic features might further increase the precision of the regression models. On the other hand, higher cognitive processes, such as social inference, may contribute to the recognition of the communicative intention (see below; Wichmann, 2000, 2002; Szameitat et al., 2010). Still, the fact that different acoustic patterns can explain the perception of different speech acts, generally leads to the assertion that prosody carries information about the speaker's intended meaning.

Emotions

Importantly, perception of intentions was not solely based on recognition of the speaker's emotion as shown by the additional regression analyses, taking valence and arousal of stimuli into account. As mentioned in the introduction, we do not exclude that communicative intentions are partly based on the emotional state of the speaker. Indeed, listeners could classify the stimuli in terms of valence and arousal (Fig. 2; Experiment 2). Nevertheless, regression analyses still explained a significant amount of variance of the speech act ratings and most speech act predictions were virtually unchanged after these affective components had been regressed out (Table 4). Only warning and naming showed a considerable decrease in the prediction rate which might be explained by their extreme positions on the arousal scale (Fig. 2). On the other hand, participants might have first identified the speaker's intention and then assigned the corresponding valence and arousal because they were asked to do so in the experiment (Experiment 2). Overall, although emotional connotations may be important for the recognition of some speech acts, our results give no reason to assume a systematic influence of emotions on the recognition of communicative intentions.

Ratings vs. 6-AFC

Finally, it is of note that the ratings replicated the results of the 6-AFC task used in Experiment 2. Importantly, ratings are a more sensitive measure than forced-choice categorization tasks because they not only allow participants to reject predefined response categories but also to flexibly adjust their responses on every (visual-analog) speech act scale. The fact that both typicality ratings in Experiment 3 and 6-AFC judgments in Experiment 2 (conducted in separate participant groups) yielded very similar results demonstrates the robustness of our findings.

General discussion

Action-theoretic views of language (Austin, 1962; Bühler, 1934; Grice, 1957; Searle, 1969) propose that speakers' intentions are the main core and driver of interpersonal communication. Yet, speakers rarely express their intentions literally in the propositional content of an utterance, raising the question of how the speaker's meaning is transmitted from sender to receiver. Here, we conceptualized intentions in terms of speech acts and provide evidence that prosody serves as an extralinguistic channel to convey intentions non-verbally. Acoustically, speakers used distinct prosodic feature configurations for different speech acts. Behaviorally, listeners were well able to differentiate these intentions from voice tone alone, even when no semantic meaning (non-words) or situational context was available (single words). Further, a direct link between acoustics and perception was demonstrated, in that acoustic features reliably (although not fully) accounted for the listeners' perception of the stimulus—even when the emotional connotation of the stimuli was controlled for.

Notably, our results were consistent across all three experiments. For example, in all measures from acoustics to perception, warning was classified with highest and criticism with lowest accuracy. Moreover, in both the stimulus-based discriminant analyses (Experiment 1) and the multiple regressions (Experiment 3), pitch rise, mean f0, as well as mean intensity and duration were the most and spectral features the least relevant cues for correct speech act recognition. This is consistent with a special role of pitch features observed in similar studies on verbal emotion (Banse & Scherer, 1996) and attitude recognition (Blanc & Dominey, 2003). Overall, the consistency of our results across experiments and participant groups lends strong support for the relevance of prosody in conveying communicative intentions.

Conventional prosodic expressions

Our results invite the assumption that speakers' intents are expressed in conventionalized prosodic forms. This view is supported (i) by the consistency of the prosodic patterns across four independent speakers for each of the six speech acts, and (ii) by the robustness of listeners' performance in identifying the expressed intentions, despite absence of contextual or semantic information. Arguably, prosodic patterns do not refer to communicative intentions as unambiguously as words refer to objects in the world. Rather we propose that they represent "communicative complexes" that connote a set of conceptually related pragmatic categories (e.g., speech acts), whose distributions of relevant acoustic cues partly overlap. This acoustic and conceptual overlap may account for the confusion of criticism and doubt in our experiments and predicts a rather loose labeling of speakers' intentions in open choice tasks, licensing our use of forced-choice task and typicality ratings (see below). Notably, our data suggest that the acoustic characteristics of these "complexes" are conventionalized to the extent that listeners can infer the relevant communicative concept by matching the perceived prosodic pattern with an internalized probabilistic distribution of acoustic cue configurations for different intentions.

Such a direct recognition of speakers' intent from prosody is reminiscent of previous work on written speech acts (Holtgraves, 2008a) suggesting that the default interpretation of illocutionary force can be based on generalized rather than particularized implicatures, i.e. can be directly understood without contextual information, similar to most idioms (e.g., to call it a day) or metaphors (e.g., He is a walking dictionary) (Glucksberg, 2003; Glucksberg, Gildea, & Bookin, 1982; Keysar, 1989). The relevance of context for the classification of speakers' prosodic intentions or attitudes has been a matter of debate for a while (Cutler, 1976; Wichmann, 2000, 2002). Some accounts posit that prosody mainly acts in a contrastive way, without conveying meaning by itself (e.g. Attardo, Eisterhold, Hay, & Poggi, 2003; Bryant & Fox Tree, 2005). By deviating from its "default", prosody is thought to motivate the listener to look for "unspoken" meanings in the utterance, i.e. to infer implicit speech actions from literal meaning by taking context information into account (Cutler & Isard, 1980; Levinson, 2013). However, as Wichmann (2002) rightly pointed out, this view requires knowledge about prosodic "defaults". We argue that this knowledge is best characterized as experience-dependent inventory of situationally distinct acoustic patterns that allows listeners to recognize broad communicative concepts based on prosody. Such a distinguished role for prosody in intention transmission is supported by the fact that these communicative concepts could be conveyed despite absence of contextual information and without knowledge of the lexical content in non-words (Experiments 2 and 3).

Note that we do not claim that context plays no role at all. Very much like lexical and syntactic processing is not based on acoustics alone but varies with context (for example in case of homophones such as "meet" vs. "meat" or ambiguous word category as in "report"; for review, see Piantadosi, Tily, & Gibson, 2012), also the prosodic recognition of speakers' intents can be shaped by context (Tanenhaus, Kurumada, & Brown, 2015). First, context predicts what interpretations are likely and may, thus, resolve perceptual ambiguity between overlapping distributions within the "communicative complex", e.g., allowing listeners to better discriminate between doubt and criticism. Second, context provides a sample of the speaker's prosodic "style" that allows listeners to flexibly adapt (even reverse) their prosodic interpretations accordingly (Tanenhaus et al., 2015). Altogether, we conclude that (paralinguistic) prosody is a signal that is able to convey a broad communicative concept on its own but becomes cognitively interlinked and specified with complementary contextual information, if available.

Prosody's initial relevance for social communication

Overall, the transfer of intentions via prosody might be a capability that forms the initial, non-linguistic foundation of interpersonal communication (Bates, Camaioni, & Volterra, 1975; Dore, 1975) that becomes gradually complemented and refined—yet not erased—by growing verbal capacities, over the course of ontogeny and perhaps even phylogeny (Oller & Griebel, 2014). For example, primate calls have been found to signal the producer's interactive stance intentionally (Schel, Townsend, Machanda, Zuberbühler, & Slocombe, 2013) via distinct acoustic structures (Crockford & Boesch, 2003; Seyfarth & Cheney, 2014), even if they lack lexical (referential) meaning (Wheeler & Fischer, 2012). Developmentally, young infants start to produce acoustically distinct prosodic patterns in the middle of their first year of life that are initially used in communicative as opposed to self-centered emotional or exploratory contexts (Papaeliou & Trevarthen, 2006; Papaeliou et al., 2002), later express specific "primitive intents" (Esteve-Gibert & Prieto, 2013; Prieto et al., 2012) and endow pointing gestures with communicative goals (Grünloh & Liszkowski, 2015). Notably, interactive prosodic patterns emerge earlier than verbal skills and become meaningful communicative instruments, most likely because parents differentiate their responses based on the acoustics of the child's vocalizations (cf. Lester et al., 1995; Oller & Griebel, 2014). The present data show that prosody continues to be indicative of speakers' intents in adulthood, despite mature verbal skills. More than that, the data suggest that the use of prosodic cues evolves further beyond infancy to express more complex intentions than those infants would ever produce (e.g., criticism or doubt). Whether speakers resort more strongly to these (early) prosodic building blocks of communication when verbal capacities may get lost or are nonexistent as in conditions of non-fluent aphasia (Barrett, Crucian, Raymer, & Heilman, 1999; Warren, Warren, Fox, & Warrington, 2003) or foreign languages is an interesting topic for future research.

Prosody in natural language use

Single-word utterances are part of our everyday life and humans start to use prosody to code for different pragmatic intentions in single words in early infancy (Dore, 1975; Prieto et al., 2012). Yet, compared to longer sentences with additional semantic information, the brevity of the present context-free stimuli may have led speakers to emphasize the relevant prosodic features. Listeners, in turn, may be more used to decode intentions in sentential contexts that often resolve ambiguities (even if ambiguities were mitigated by the 6-AFC task and typicality ratings in the present study). Future studies can help to generalize our results by using a wider set of recordings (as suggested by Banse & Scherer, 1996), for example, including sentence-level stimuli, more variable tokens (i.e. more words/sentences), more speech acts, and more speakers.

Apart from that, another point of discussion is in how far prosodies produced in the sound lab using fictional scenarios correspond to prosodies produced in natural conversations. Although a direct empirical investigation is still pending, there are several reasons that grant the ecological validity of our sound stimuli. First, cues for expressing intentions are typically produced voluntarily during an interaction. Therefore, they have a posed character by nature and may not suffer from artificial recording situations to the extent as emotions do (Jürgens, Grass, Drolet, & Fischer, 2015; Jürgens et al., 2011). Second, our speakers—although trained in producing clear and artifact-free speech—were non-actors. Hence, they relied on their everyday speech experience to express the intention in a way they would naturally do to be understood by an interlocutor. Last, studies on non-prosodic cues for speech acts (Bucciarelli et al., 2003; Reeder, 1980) and voluntary vocal expressions of social affect (Rilliard, Shochi, Martin, Erickson, & Auberge, 2009) suggest that cues for expressing intentions are not innate but culturally learned. On this assumption, the fact that our speakers and listeners used and understood the specific prosodic cues suggests that these cues must occur in natural conversations.

Future research on intentional prosody

An interesting question with regard to speaker's intent in natural communication is, then, how prosodic cues are weighed and cognitively interlinked with other paralinguistic cues such as facial expressions. Notably, the latter have been shown to serve explicit interpersonal functions that reach beyond the inadvertent display of basic emotions (Ekman, 1992), for example when (voluntarily) communicating *social motives* (e.g., in case of compassion or empathy for pain) (Fridlund, 1994; Parkinson, 2005). Concerning audio–visual integration, recent motion-capture and neuroimaging studies revealed interactions between linguistic/emotional prosody and facial expressions, in speakers (Cvejic, Kim, & Davis, 2012; Kitamura, Guellaï, & Kim, 2014) and in listeners, respectively (Brück, Kreifelts, & Wildgruber, 2011; Watson et al., 2014). Yet, whether and how prosody and facial cues are fused in the transfer of speaker's meaning is currently not known and an interesting topic for future research.

Another point that deserves further examination is our observation that acoustic information predicted participant's speech act recognition successfully, yet not fully. This raises the interesting hypothesis that the comprehension of speaker's intentions from prosody relies on a weighted contribution of auditory-prosodic and other, socio-cognitive processes whose exact nature and ways of interaction still need to be clarified. On the socio-cognitive side, recent neuroimaging work lends initial evidence for inferential processes, i.e. involving theory of mind areas, during the comprehension of speech acts (Egorova, Shtyrov, & Pulvermüller, 2015; Egorova et al., 2014) and speaker meaning (Bašnáková, Weber, Petersson, van Berkum, & Hagoort, 2014; Jang et al., 2013), as well as motor system involvement during the processing of directive speech acts (Egorova et al., 2014) and indirect requests (van Ackeren, Casasanto, Bekkering, Hagoort, & Rueschemeyer, 2012). Yet, none of these studies involved prosody, leaving the fundamental question unresolved how prosody potentially interlinks with these socio-cognitive systems. Future neurocognitive investigations with the present stimuli may help to elucidate this question and are currently underway.

Conclusion

Speakers rarely code their intentions in the lexical content of an utterance. Yet, listeners easily recognize the speaker's communicative goals. The present study shows that conversationalists are able to use prosody as extralinguistic cue to specify communicative intentions—an early capacity that complements adults' mature verbal abilities. Interlocutors produce and understand prosodic cues independently of the semantic meaning, contextual information, and emotional coloring of the utterance. These results argue in favor of conventionalized acoustic feature configurations that connote communicative concepts, although their acoustic and conceptual distributions may partly overlap. The present study leads toward future research on the interaction between auditory-prosodic cues, conversation context, and socio-cognitive processes serving the transfer of speaker meaning as the foundation of successful interpersonal communication.

Acknowledgments

We thank Prof. Dr. J.D. Jescheniak and Prof. Dr. K. von Kriegstein for their helpful feedback on this project. The study was funded by the Otto Hahn award of the Max Planck Society to DS.

Appendix A

Situation descriptions presented to the speakers for speech act recordings (English translations). All speakers read the scenarios and were asked to place themselves in the interpersonal situation, for example, by uttering the example sentences (in bold), before uttering the single words (or non-words) with the corresponding prosody. Scenarios contained either words or non-words (in square brackets below).

Criticism

Your colleague Tom and you will present your first big job in an important meeting this afternoon. Therefore, you are extremely nervous and do not want to disappoint your boss. You sit at your desk and go through the presentation one more time. Suddenly, there is a knock on the door and Tom peeks into your office. He asks you whether you would like to join him for a beer [diem] in the bar [dahm], although he knows how important the upcoming meeting is. You think it would be best for him to prepare the joint talk and ask disapprovingly:

BIER [DIEM]: **"(Are you serious? A) beer [diem], (now)?"**
BAR [DAHM]: **"(Are you serious? The) bar [dahm], (now)?"**

Doubt

You arrive home after a hard day at work and it is quite late. Your mobile rings when you have just hung your coat up. It is your friend, Eva, and she suggests having a beer [diem] at the bar [dahm]. You would actually like to meet her as you have not seen her for a long time, but are very tired and need to leave for work early the next day. You do not know whether this is a good idea. Therefore, you ask doubtfully:

BEER [DIEM]: **"(A) beer [diem], (now)?"**
BAR [DAHM]: **"(The) bar [dahm], (now)?"**

Naming

Please say the words: beer/bar/diem/dahm with a neutral intonation, for example, as in the sentence:

"(I'm going to have a) beer [diem] (tonight)."
"(I'm going to a) bar [dahm] (tonight)."

Suggestion

It is Thursday evening and you have almost finished your work. You achieved a lot today and are satisfied with your work. You really deserve to go to the bar [dahm] for a beer [diem] now. You think you can perhaps convince your colleague, Anne, to join as you sometimes go to your favorite bar together after work. In pleasant anticipation of a nice evening, you peek into Anne's office and ask invitingly:

BEER [DIEM]: **"(Are you up for a) beer [diem]?"**
BAR [DAHM]: **"(Do you want to go to a) bar [dahm]?"**

Warning

You invited a friend to have a beer at your apartment. He talks excitedly about his last football match and vividly tries to imitate one of his maneuvers. He spins around wildly, back and forth, left and right, and you start getting worried about your furniture. He suddenly starts running and does not see your mini bar [dahm] where he put his glass of beer [diem]. You try to warn him:

BEER [DIEM]: **"(Watch out, your) beer [diem]!"**
BAR [DAHM]: **"(Watch out, the) bar [dahm]!"**

Wish

It is a hot summer day and you descend after a hard, but wonderful mountain hike. After all those kilometers and the great view, you are pleasantly exhausted, hungry, and thirsty—the hotel is almost within sight. You only have one thought on your mind: You would like a nice cool beer [diem] at the hotel bar [dahm] to make this day truly perfect. You say longingly:

BEER [DIEM]: **"(Now for a) beer [diem]!"**
BAR [DAHM]: **"(Now to the) bar [dahm]!"**

Appendix B

Definitions of speech acts presented to participants before the behavioral tests (English translations).

Criticism

The speaker, a friend of yours, is disapprovingly expressing criticism, for example, about one of your suggestions.

Doubt

The speaker is deliberately expressing doubt, for example about whether to accept a proposal you made.

Naming

The speaker is saying something for no specific purpose, for example, to name an object.

Suggestion

The speaker is invitingly suggesting something, for example, to undertake something together.

Warning

The speaker is warning you of a possible accident, for example, not to fall over an object.

Wish

The speaker is longingly expressing a wish for something, for example, a relaxing evening after a successful working day.

Appendix C

See Tables C1–C3.

Table C1
Mean acoustic features per speech act category.

Speech act	Acoustic feature						
	Number of voiced frames	Mean f0 (Hz)	Offset–onset f0 (Hz)	Mean intensity (dB)	Mean HNR (dB)	Spectral center of gravity (Hz)	*SD* spectrum (Hz)
Words							
Criticism	450.1 ± 61.1	230.7 ± 48.4	81.5 ± 91.6	65.1 ± 3.9	12.1 ± 3.2	712.1 ± 263.8	728.6 ± 243.4
Doubt	482.8 ± 68.7	188.9 ± 41.6	72.6 ± 30.9	57.2 ± 3.6	14.2 ± 3.4	511.4 ± 186.0	713.6 ± 204.7
Naming	341.1 ± 64.5	13.3 ± 42.6	−51.6 ± 29.0	56.9 ± 4.2	13.3 ± 2.3	587.7 ± 248.1	554.4 ± 139.1
Suggestion	320.0 ± 56.4	206.1 ± 37.2	184.7 ± 55.2	63.3 ± 2.0	13.2 ± 2.7	617.1 ± 224.0	611.7 ± 136.0
Warning	428.4 ± 97.9	268.5 ± 49.4	−122.7 ± 27.0	71.8 ± 2.3	13.9 ± 2.8	897.1 ± 209.3	762.2 ± 244.4
Wish	485.7 ± 62.9	148.6 ± 39.5	−61.7 ± 16.4	59.1 ± 3.1	12.7 ± 2.1	591.1 ± 238.2	781.1 ± 319.6
Average	418.0 ± 95.3	196.9 ± 62.3	17.2 ± 115.5	62.2 ± 6.2	13.2 ± 2.9	652.8 ± 259.8	652.8 ± 237.0
Non-words							
Criticism	473.0 ± 81.1	250.1 ± 60.7	118.9 ± 112.6	64.8 ± 3.6	15.5 ± 4.7	645.5 ± 279.2	806.2 ± 227.4
Doubt	510.7 ± 57.5	185.8 ± 47.5	70.9 ± 30.8	57.0 ± 4.1	18.9 ± 4.1	386.8 ± 118.9	641.9 ± 230.0
Naming	427.3 ± 72.9	134.1 ± 41.7	−56.8 ± 28.4	55.2 ± 4.2	17.2 ± 2.8	504.2 ± 235.0	597.5 ± 174.1
Suggestion	342.0 ± 50.8	217.4 ± 41.0	213.2 ± 49.5	62.2 ± 3.1	16.1 ± 3.6	455.0 ± 169.0	567.9 ± 154.4
Warning	474.2 ± 124.9	280.5 ± 46.7	−125.3 ± 23.3	71.4 ± 1.6	18.1 ± 2.6	755.1 ± 288.8	713.7 ± 184.0
Wish	560.8 ± 87.1	142.3 ± 38.0	−53.2 ± 19.0	57.9 ± 3.5	15.9 ± 3.9	494.5 ± 188.3	760.9 ± 205.4
Average	464.7 ± 106.8	201.7 ± 70.7	27.9 ± 128.7	61.4 ± 6.5	17.0 ± 3.9	540.2 ± 252.5	681.4 ± 214.6

Values depict mean ± *SD*. HNR = harmonics-to-noise ratio; *SD* = standard deviation. All values were extracted using PRAAT 5.3.01 (http://www.praat.org).

Table C2
Statistical comparison of acoustic features between words and non-words.

Acoustic parameter	Speech act ratings											
	Criticism		Doubt		Naming		Suggestion		Warning		Wish	
	t(6)	*p*	*t*(6)	*p*	*t*(6)	*p*	*t*(6)	*p*	*t*(6)	*p*	*t*(6)	*p*
Voiced frames	−0.453	.666	−1.376	.218	−1.831	.117	−0.710	.504	−0.512	.627	−1.679	.144
Mean f0	−0.427	.684	−0.027	.980	0.124	.906	−0.191	.854	−0.272	.795	0.251	.810
Offset–onset f0	−0.299	.775	0.277	.791	0.630	.552	−0.618	.559	−0.198	.849	−1.652	.150
Mean intensity	0.159	.879	−0.191	.855	0.605	.567	0.837	.435	0.105	.920	0.903	.401
Mean HNR	−1.735	.134	**−7.098**	**.000**	**−2.749**	**.033**	−2.317	.060	**−2.659**	**.038**	−2.309	.060
Center of gravity	0.605	.567	**2.619**	**.040**	1.193	.278	**4.292**	**.005**	2.011	.091	**2.562**	**.043**
SD spectrum	−0.962	.373	0.611	.564	−0.415	.693	0.689	.516	0.443	.673	−0.008	.994

Acoustic features of words and non-words were compared with paired *t*-tests for each speech act category (columns). Significant results ($p < .05$) are marked in bold. f0 = fundamental frequency; HNR = harmonics-to-noise ratio; *SD* = standard deviation.

Table C3
Results of the discriminant analyses (Experiment 1).

	Words	Non-words
Function 1		
Offset–onset f0	0.881	0.836
Function 2		
Mean intensity	0.721	0.716
Mean f0	0.467	0.538
Function 3		
Voiced Frames	0.768	0.708

Within-group correlations between acoustic measures and standardized canonical discriminant functions. Table includes values of the first three functions above a threshold of $r = 0.4$.

Appendix D. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2016.01.001.

References

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor – International Journal of Humor Research, 16*, 243–260.

Austin, J. L. (1962). *How to do things with words*. Cambridge: Harvard University Press.

Bach, K. (1994). Conversational impliciture. *Mind & Language, 9*, 124–162.

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

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication, 46*, 252–267.

Barrett, M. D., Crucian, G. P., Raymer, A. M., & Heilman, K. M. (1999). Spared comprehension of emotional prosody in a patient with global aphasia. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 12*, 117–120.

Bašnáková, J., Weber, K., Petersson, K. M., van Berkum, J., & Hagoort, P. (2014). Beyond the language given: The neural correlates of inferring speaker meaning. *Cerebral Cortex, 24*, 2572–2578.

Bates, E., Camaioni, L., & Volterra, V. (1975). The acquisition of performatives prior to speech. *Merrill-Palmer Quarterly, 21*, 205–226.

Blanc, J., & Dominey, P. (2003). Identification of prosodic attitudes by a temporal recurrent network. *Cognitive Brain Research, 17*, 693–699.

Boersma, P., & Weenink, D. (2014). *Praat: Doing phonetics by computer [computer program]. Version 5.3.80*. <http://www.praat.org/> Retrieved 29.06.14.

Bolinger, D. (1986). *Intonation and its parts: Melody in spoken English*. Stanford: Stanford University Press.

Bradley, M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manakin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*, 49–59.

Brück, C., Kreifelts, B., & Wildgruber, D. (2011). Emotional voices in context: A neurobiological model of multimodal affective information processing. *Physics of Life Reviews, 8*, 383–403.

Bryant, G. A., & Fox Tree, J. E. (2005). Is there an ironic tone of voice? *Language and Speech, 48*, 257–277.

Bucciarelli, M., Colle, L., & Bara, B. G. (2003). How children comprehend speech acts and communicative gestures. *Journal of Pragmatics, 35*, 207–241.

Bühler, K. (1934). *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Jena: Gustav Fischer.

Carlson, K., Frazier, L., & Clifton, C. J. (2009). How prosody constrains comprehension: A limited effect of prosodic packaging. *Lingua, 119*, 1066–1082.

Clark, H. H., & Carlson, T. B. (1981). Context for comprehension. In J. Ling & A. Baddeley (Eds.), *Attention and performance IX* (pp. 313–330). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Crockford, C., & Boesch, C. (2003). Context-specific calls in wild chimpanzees, *Pan troglodytes verus*: Analysis of barks. *Animal Behaviour, 66*, 115–125.

Cutler, A. (1976). *The context-dependence of "intonational meanings"*, pp. 104–115.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*, 141–201.

Cutler, A., & Isard, S. D. (1980). The production of prosody. In B. Butterworth (Ed.). *Language production: Speech and talk* (Vol. 1, pp. 245–270). Academic Press.

Cvejic, E., Kim, J., & Davis, C. (2012). Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody. *Cognition, 122*, 442–453.

Di Cesare, G., Di Dio, C., Marchi, M., & Rizzolatti, G. (2015). Expressing our internal states and understanding those of others. *Proceedings of the National Academy of Sciences, 112*, 10331–10335.

Dore, J. (1975). Holophrases, speech acts and language universals. *Journal of Child Language, 2*, 21–40.

Egorova, N., Pulvermüller, F., & Shtyrov, Y. (2014). Neural dynamics of speech act comprehension: An MEG study of naming and requesting. *Brain Topography, 27*, 375–392.

Egorova, N., Shtyrov, Y., & Pulvermüller, F. (2013). Early and parallel processing of pragmatic and semantic information in speech acts: Neurophysiological evidence. *Frontiers in Human Neuroscience, 7*, 1–13.

Egorova, N., Shtyrov, Y., & Pulvermüller, F. (2015). Brain basis of communicative actions in language. *NeuroImage, 125*, 857–867.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*.

Enrici, I., Adenzato, M., Cappa, S., Bara, B. G., & Tettamanti, M. (2011). Intention processing in communication: A common brain network for language and gestures. *Journal of Cognitive Neuroscience, 23*, 2415–2431.

Esteve-Gibert, N., & Prieto, P. (2013). Prosody signals the emergence of intentional communication in the first year of life: Evidence from Catalan-babbling infants. *Journal of Child Language*.

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin, 97*, 412–429.

Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. Academic Press.

Frith, C. (2009). Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 364*, 3453–3458.

Furrow, D., Podrouzek, W., & Moore, C. (1990). The acoustical analysis of children's use of prosody in assertive and directive contexts. *First Language, 10*, 37–49.

Gisladottir, R. S., Chwilla, D. J., & Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PLoS ONE, 10*, 1–24.

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in Cognitive Sciences, 7*, 92–96.

Glucksberg, S., Gildea, P., & Bookin, H. B. (1982). On understanding nonliteral speech: Can people ignore metaphors? *Journal of Verbal Learning and Verbal Behavior, 21*, 85–98.

Grice, H. P. (1957). Meaning. *The Philosophical Review, 66*, 377–388.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics: 3. Speech acts* (pp. 41–58). New York: Academic Press.

Grünloh, T., & Liszkowski, U. (2015). Prelinguistic vocalizations distinguish pointing acts. *Journal of Child Language, 49*, 1–48.

Holtgraves, T. (2002). *Language as social action*. London: Lawrence Erlbaum Associates.

Holtgraves, T. (2005). The production and perception of implicit performatives. *Journal of Pragmatics, 37*, 2024–2043.

Holtgraves, T. (2008a). Automatic intention recognition in conversation processing. *Journal of Memory and Language, 58*, 627–645.

Holtgraves, T. (2008b). Conversation, speech acts, and memory. *Memory & Cognition, 36*, 361–374.

Jang, G., Yoon, S. A., Lee, S. E., Park, H., Kim, J., Ko, J. H., & Park, H. J. (2013). Everyday conversation requires cognitive inference: Neural bases of comprehending implicated meanings in conversations. *NeuroImage, 81*, 61–72.

Jiang, X., & Pell, M. D. (2015). On how the brain decodes vocal cues about speaker confidence. *Cortex, 66*, 9–34.

Jürgens, R., Grass, A., Drolet, M., & Fischer, J. (2015). Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected. *Journal of Nonverbal Behavior*, 195–214.

Jürgens, R., Hammerschmidt, K., & Fischer, J. (2011). Authentic and play-acted vocal emotion expressions reveal acoustic differences. *Frontiers in Psychology, 2*, 1–11.

Keysar, B. (1989). On the functional equivalence of literal and metaphorical interpretations in discourse. *Journal of Memory and Language, 28*, 375–385.

Kitamura, C., Guellaï, B., & Kim, J. (2014). Motherese by eye and ear: Infants perceive visual prosody in point-line displays of talking heads. *PLoS ONE, 9*, e111467.

Lester, B. M., Boukydis, C. F., Zachariah Garcia-Coll, C. T., Peucker, M., McGrath, M. M., Vohr, B. R., ... Oh, W. (1995). Developmental outcome as a function of the goodness of fit between the infant's cry characteristics and the mother's perception of her infant's cry. *Pediatrics, 95*, 516–521.

Levinson, S. C. (2006). On the "human interaction engine". In N. J. Enfield & S. C. Levinson (Eds.), *Roots of human sociality: Culture, cognition, and interaction* (pp. 39–69). Oxford: Berg.

Levinson, S. C. (2013). Action formation and ascription. In T. Stivers & J. Sidnell (Eds.), *The handbook of conversation analysis* (pp. 103–130). Wiley-Blackwell.

Liu, S. (2011). An experimental study of the classification and recognition of Chinese speech acts. *Journal of Pragmatics, 43*, 1801–1817.

Marcos, H. (1987). Communicative functions of pitch range and pitch direction in infants. *Journal of Child Language, 14*, 255.

Mead, G. H. (1934). In C. W. Morris (Ed.), *Mind, self, and society: From the standpoint of a social behaviorist*. University of Chicago Press.

Monetta, L., Cheang, H. S., & Pell, M. D. (2008). Understanding speaker attitudes from prosody by adults with Parkinson's disease. *Journal of Neuropsychology, 2*, 415–430.

Morlec, Y., Bailly, G., & Aubergé, V. (2001). Generating prosodic attitudes in French: Data, model and evaluation. *Speech Communication, 33*, 357–371.

Oller, D. K., & Griebel, U. (2014). On quantitative comparative research in communication and language evolution. *Biological Theory, 9*, 296–308.

Papaeliou, C., Minadakis, G., & Cavouras, D. (2002). Acoustic patterns of infant vocalizations expressing emotions and communicative functions. *Journal of Speech, Language, and Hearing Research, 45*, 311–317.

Papaeliou, C., & Trevarthen, C. (2006). Prelinguistic pitch patterns expressing "communication" and "apprehension". *Journal of Child Language, 33*, 163–178.

Parkinson, B. (2005). Do facial movements express emotions or communicate motives? *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc, 9*, 278–311.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition, 122*, 280–291.

Prieto, P., Estrella, A., Thorson, J., & Vanrell, M. D. M. (2012). Is prosodic development correlated with grammatical and lexical development? Evidence from emerging intonation in Catalan and Spanish. *Journal of Child Language, 39*, 221–257.

Reeder, K. (1980). The emergence of illocutionary skills. *Journal of Child Language, 7*, 13–28.

Remmington, N. a., Fabrigar, L. R., & Visser, P. S. (2000). Reexamining the circumplex model of affect. *Journal of Personality and Social Psychology, 79*, 286–300.

Rigoulot, S., Fish, K., & Pell, M. D. (2014). Neural correlates of inferring speaker sincerity from white lies: An event-related potential source localization study. *Brain Research, 1565*, 48–62.

Rilliard, A., Shochi, T., Martin, J.-C., Erickson, D., & Auberge, V. (2009). Multimodal indices to Japanese and French prosodically expressed social affects. *Language and Speech, 52*, 223–243.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*.

Sammler, D., Grosbras, M.-H., Anwander, A., Bestelmeyer, P. E. G., & Belin, P. (2015). Dorsal and ventral pathways for prosody. *Current Biology*, 1–7.

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology, 63*, 2251–2272.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 2408–2412.

Schel, A. M., Townsend, S. W., Machanda, Z., Zuberbühler, K., & Slocombe, K. E. (2013). Chimpanzee alarm call production meets key criteria for intentionality. *PLoS ONE, 8*.

Schneider, K., Lintfert, B., Dogil, G., & Möbius, B. (2006). Phonetic grounding of prosodic categories. In *Methods in empirical prosody research* (pp. 335–362). Berlin: De Gruyter.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

Searle, J. R., & Vanderveken, D. (1985). *Foundation of illocutionary logic.* Cambridge University Press.

Seyfarth, R. M., & Cheney, D. L. (2014). The evolution of language from social cognition. *Current Opinion in Neurobiology, 28*, 5–9.

Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion (Washington, D.C.), 9*, 838–846.

Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech, 46*, 1–22.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy, 25*, 701–721.

Szameitat, D. P., Alter, K., Szameitat, A. J., Darwin, C. J., Wildgruber, D., Dietrich, S., & Sterr, A. (2009). Differentiation of emotions in laughter at the behavioral level. *Emotion (Washington, D.C.), 9*, 397–405.

Szameitat, D. P., Alter, K., Szameitat, A. J., Wildgruber, D., Sterr, A., & Darwin, C. J. (2009). Acoustic profiles of distinct emotional expressions in laughter. *The Journal of the Acoustical Society of America, 126*, 354–366.

Szameitat, D. P., Kreifelts, B., Alter, K., Szameitat, A. J., Sterr, A., Grodd, W., & Wildgruber, D. (2010). It is not always tickling: Distinct cerebral responses during perception of different laughter types. *NeuroImage, 53*, 1264–1271.

Tanenhaus, M. K., Kurumada, C., & Brown, M. (2015). Prosody and intention recognition. In L. Frazier & E. Gibson (Eds.), *Explicit and implicit prosody in sentence processing* (pp. 99–118). Springer.

Tomasello, M. (2005). Intention reading and imitative learning. In S. Hurley & N. Chater (Eds.), *Perspectives on imitation: From neuroscience to social science* (pp. 133–148). MA: MIT Press.

Uldall, E. (1960). Attitudinal meanings conveyed by intonation contours. *Language and Speech, 3*, 223–234.

van Ackeren, M. J., Casasanto, D., Bekkering, H., Hagoort, P., & Rueschemeyer, S.-A. (2012). Pragmatics in action: Indirect requests engage theory of mind areas and the cortical motor network. *Journal of Cognitive Neuroscience, 24*, 2237–2247.

Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes, 25*, 905–945.

Warren, P. (1999). Prosody and language processing. In *Language processing* (pp. 155–188). Psychology Press Ltd.

Warren, J. D., Warren, J. E., Fox, N. C., & Warrington, E. K. (2003). Nothing to say, something to sing: Primary progressive dynamic aphasia. *Neurocase, 9*, 140–155.

Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal adaptation in right posterior superior temporal sulcus during face-voice emotional integration. *Journal of Neuroscience, 34*, 6813–6821.

Wheeler, B. C., & Fischer, J. (2012). Functionally referential signals: A promising paradigm whose time has passed. *Evolutionary Anthropology: Issues, News, and Reviews, 21*, 195–205.

Wichmann, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, pp. 143–148.

Wichmann, A. (2002). Attitudinal intonation and the inferential process. *Proceedings of Speech Prosody Conference*, Aix-en-Provence.

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge University Press.

Wittgenstein, L. (1953). Philosophische Untersuchungen. *Suhrkamp*.

Wundt, W. (1896). *Grundriss der Psychologie*. Leipzig: Engelmann.