What's in a word: Sounding sarcastic in British English

Aoju Chen
Utrecht University
aoju.chen@uu.nl

Lou Boves
Radboud University Nijmegen
l.boves@ru.nl

Using a simulated telephone conversation task, we elicited sarcastic production in different utterance types (i.e. declaratives, tag questions and *wh*-exclamatives from native speakers of the southern variety of British English. Unlike previous studies which focus on static prosodic measurements at the utterance level (e.g. mean pitch, pitch span, intensity), we examined both static prosodic measurements and continuous changes in contour shape in the semantically most important words (or key words) of the sarcastic utterances and their counterparts in the sincere utterances. To this end, we adopted Functional Data Analysis to model pitch variation as contours and represent the contours as continuous function in statistical analysis. We found that sarcasm and sincerity are prosodically distinguishable in the key words alone. The key words in sarcastic utterances are realised with a longer duration and a flatter fall than their counterparts in sincere utterances regardless of utterance type and speaker gender. These results are compatible with previous reports on the use of a smaller pitch span and a slower speech rate in sarcastic utterances than in sincere utterances in North American English. We also observed notable differences in the use of minimum pitch, maximum pitch and contour shape in different utterance types and in the use of mean pitch and duration by male and female speakers. Additionally, we found that the prosody of the key words in sarcastic utterances and their counterparts in sincere utterances has yielded useful predictors for the presence (or absence) of sarcasm in an utterance. Together, our results lend direct support to a key-word–based approach. However, the prosodic predictors included in our analysis alone can achieve only an accuracy of 70.4%, suggesting a need to examine additional prosodic parameters and prosody beyond the key words.

1 Introduction

Sarcasm is a common phenomenon in interactive communication. It refers to the processes in which speakers use words to express an intention opposite to the literal meaning of those words (Kreuz & Glucksberg 1989, Capelli, Nakagawa & Madden 1990, Anolli, Ciceri &

Journal of the International Phonetic Association (2018) 48/1
doi:10.1017/S0025100318000038

Infantino 2002).[1] Speakers can use sarcasm to either emphasise a criticism or soften a critical intent, and to harmonise the effect of a praise (Knox 1961; Brown & Levinson 1987; Kreuz & Glucksberg 1989; Oring 1994; Jorgensen 1996; Gibbs 1999; Clift 1999; Anolli, Ciceri & Infantino 2000, 2002; Kreuz 2000; Rockwell 2007). The use of sarcasm presents itself as an interesting challenge for developing spoken dialogue systems due to the complexity in its semantic nature and its expression (e.g. Tepperman, Traum & Narayanan 2006). A thorough understanding of how speakers express sarcasm is thus not only desirable from a theoretical perspective but also necessary for achieving accurate recognition of speaker intention and generating adequate responses in real time in highly-interactive dialogue systems (Rakov & Rosenberg 2013, Ward & DeVault 2017).

Speakers use both verbal means (e.g. syntactic, lexical, and vocal cues) and non-verbal means (e.g. facial expressions) to convey sarcasm (Rockwell 2001, Attardo et al. 2003). The importance of vocal cues for the expression of sarcasm has long been recognised in theories of verbal irony (e.g. Clark & Gerrig 1984, Sperber 1984), but how speakers use vocal cues has been empirically investigated only in recent years. Some researchers have argued that there are universal vocal cues for sarcasm, including a slower speaking rate, greater intensity, nasalisation, monotony, and extreme pitch levels (Rockwell 2000, 2007; Attardo et al. 2003). However, production studies examining a range of prosodic parameters in different languages have not only shown similarities but also differences in the use of prosody in expressing sarcasm across languages (e.g. Anolli et al. 2002 on Italian; Cheang & Pell 2009 on Cantonese; Loevenbruck et al. 2013 and González-Fuente, Prieto & Noveck 2016 on French; Rao 2013 on Mexican Spanish; Niebuhr 2014 on German; González-Fuente, Escandell-Vidal & Prieto 2015 on Catalan). For example, sarcastic utterances are realised with a larger pitch span (difference between the highest and lowest pitch) in Italian (Anolli et al. 2002) but a smaller pitch span in German and Mexican Spanish, compared to neutral or sincere utterances.

In the current study, we aim to obtain a deeper understanding of prosodic realisation of sarcasm in English by studying both what the prosodic differences are between sarcasm and sincerity and which prosodic and non-prosodic factors can predict the presence (or absence) of sarcasm. To this end, we take a novel linguistically-motivated approach to the locus of prosodic characteristics of sarcasm and address major methodological limitations in previous analysis. In the remaining part of this section, we first briefly review main findings from earlier research on prosodic realisation of sarcasm in English, then present our linguistically-motivated approach, and finally describe the methodological limitations to be tackled.

1.1 Past work on prosodic realisation of sarcasm in English

Previous production studies of prosodic realisation of sarcasm in English were exclusively concerned with North American English. For example, Rockwell (2000) examined utterances read out by professional speakers (e.g. radio announcers, actors) and experienced non-professional speakers in different contexts in American English and found that the speakers realised the utterances in a sarcastic context with a slower speech rate, lower mean pitch and greater intensity than in a neutral context. Rakov & Rosenberg (2013) were concerned with another type of read speech, i.e. acted speech by voice actors reading from a script. They analysed pitch span of each utterance and modelled the pitch and intensity contours both across the whole utterance and in each word of an utterance using Legendre polynomials (Dumouchel et al. 2009). They found that sarcasm and sincerity in acted speech could be automatically detected with about 82% accuracy on the basis of combined information on pitch span and shape of pitch and intensity contours across the utterance and in each word.

[1] The term 'sarcasm' is sometimes used interchangeably with the term 'irony' (Kreuz 2000), although there is no consensus on the relationship between them in the literature (Muecke 1980, Kreuz & Gluckberg 1989, Lee & Katz 1998, Kreuz 2000). In this study, we consider the two terms synonymous.

This finding suggests that sarcastic speech might differ from sincere speech in the shape of pitch and intensity contours as well as in pitch span. The authors observed that sarcastic speech had a smaller pitch span, fewer instances of falling pitch contours and shallowly rising pitch contours but more instances of shallowly rising intensity contours than sincere speech did.

Different from Rockwell (2000) and Rakov & Rosenberg (2013), Rockwell (2007) studied sarcastic prosody in spontaneous dialogues about pre-selected questions between familiar speakers in American English. Analysing only the utterances that were identified to convey sarcastic messages by the speakers themselves, Rockwell found that the sarcastic utterances were spoken with a higher mean pitch, larger pitch span, and longer duration, compared to the non-sarcastic utterances. The longer duration suggests a slower speech rate in sarcastic utterances in spontaneous speech, similar to what was found on read speech, whereas the use of a higher mean pitch and a larger pitch span was not found in sarcastic utterances in read speech (Rockwell 2000, Rakov & Rosenberg 2013). However, because the sarcastic utterances in Rockwell (2007) were lexically different from the non-sarcastic utterances, it casts doubt on the validity of the reported results based on direct comparisons between the sarcastic and non-sarcastic utterances. The differences in the results between these studies thus do not necessarily suggest different uses of prosody in sarcastic expression between read speech and spontaneous speech in American English. Bryant (2010) also examined the prosody in sarcastic utterances produced in spontaneous speech in American English. He analysed utterance-level prosodic changes from two utterances preceding a sarcastic utterance to the sarcastic utterance and found that only speech rate differed significantly between the sarcastic utterance and its preceding utterances. That is, speech rate was lower in the sarcastic utterance than in the preceding utterances, in line with the result on speech rate reported by Rockwell (2000, 2007).

Finally, Cheang & Pell (2008) examined utterances read out in different contexts by native speakers of Canadian English. Analysing only the utterances that were perceived to sound sarcastic by other native speakers of Canadian English, they found that the utterances were consistently realised with a lower mean pitch, less within-utterance pitch variation, and a slower speech rate (albeit only in short utterances like 'Is that so') in the sarcastic context than in the sincere and humour contexts. The result on speech rate was compatible with the finding reported for read speech (Rockwell 2000, 2007) and spontaneous speech (Bryant 2010) in American English; the result on mean pitch was similar to the finding reported in read speech in American English (Rockwell 2000). There thus appears to be more similarities than differences in prosodic realisation of sarcasm in the two varieties of North American English.

1.2 The current approach

Notably, the studies reviewed above have all taken the whole utterance as the target of analysis, i.e. analysing prosodic variation across the utterance or word-by-word, similar to the studies on sarcastic prosody in other languages (see González-Fuente et al. 2016 for an exception). Typical prosodic measurements include minimum pitch, maximum pitch, mean pitch, pitch span and speech rate of the utterance. While such a holistic approach has proved to be useful in this line of research, it is not entirely clear what the underlying linguistic or methodological motivation is. In this study, we focus on the word that is semantically critical to the expression of sarcasm in an utterance (hereafter key word) and examine whether sarcasm is prosodically distinguishable from sincerity in the key words alone. For example, in the utterance 'She's a healthy lady' said as a sarcastic response to a friend's remark about his aunt ('My aunt smokes a pack a day'), the real message of the speaker is that the friend's aunt is an unhealthy lady. The word *healthy* thus holds the key to the sarcastic interpretation of the utterance, different from the other words in the same utterance. The key word is identifiable even when multiple words can play a role in the expression of sarcasm in an utterance. Take for example the

utterance 'The service is really good here' said as a sarcastic response to the poor service in a restaurant. Although both *really* and *good* contribute to the expression of sarcasm, *good* is the key word, because the word *really* is dispensable but the opposite meeting of the word *good* is the real message.

The significance of the key word in a sarcastic utterance is comparable to that of the word in focus in an utterance. Focus is an information structural category, refers to the predication on a topic in an utterance, and typically contains new information to the receiver (Lambrecht 1994, Vallduví & Engdahl 1996). The same utterance can be a felicitous response in different contexts, which subject the same word to different focus types or different words to the scope of focus. For example, the utterance 'She's a healthy lady' can be said as a response to the question *What do you know about her?*, the question *Is she weak?*, or the question *She's a healthy what?*. In the first two cases, the word *healthy* conveys new information sought by the questioner. But the status of the word *healthy* is different. In the first case, the whole verb phrase (*is a healthy lady*) is in focus (i.e. broad focus); the word *healthy* is part of the broad focus. In the second case, only the word *healthy* conveys new and contrastive information; the word *healthy* is thus in narrow contrastive focus. In the third case, the word *lady* is focal but the word *healthy* is not. The information structural difference between the three renditions of the utterance is primarily expressed in the prosody of the word *healthy*, although the post-focus word *lady* also undergoes some prosodic changes in the second case (e.g. Chen 2012). In the light of the parallel between the key word in a sarcastic utterance and the focal word in an utterance, we hypothesise that prosodic differences between sarcasm and sincerity can be found in the key words alone in British English. Preliminary evidence has been reported in a recent study on vocal cues to sarcasm in French. González-Fuente et al. (2016) examined prosodic differences between the sarcastic and sincere renditions of six declaratives with the key word in utterance-final position, which were read aloud by ten female natives speakers of French. They found similar prosodic differences between the two renditions of utterances at the utterance- and word-levels.

1.3 Open methodological issues

Previous studies of sarcastic prosody in English (and other languages) have frequently represented pitch as static values (e.g. mean pitch, pitch-maximum, pitch-minimum, pitch span) in statistical analyses. This method stands in sharp contrast to the fact that pitch is a continuous variable and changes over time in an utterance, resulting in variation in the shape of the pitch contour. Our understanding of the role of contour shape in sarcastic expression is consequently still very limited. Traditionally, researchers transcribe the shape of the pitch contour in individual words of an utterance using two discrete tones (i.e. high and low tones) and their modifications (i.e. downstep, upstep), following notations proposed within the autosegmental–metrical framework (Pierrehumbert 1980, Ladd 1996). Such an approach can generate valuable insights into the role of contour shape in the expression of sarcasm (González-Fuente et al. 2016) and other attributes. But it requires substantial knowledge of intonational phonology and training in the annotators and labour-intensive inter-rater agreement experiments to assess the reliability of the resulting annotation (Grabe, Kochanski & Coleman 2007). Also, it is questionable whether the discrete representation of the pitch contour in individual words can sufficiently capture subtle variation in the shape of pitch contour involved in the expression of affect and attitude such as sarcasm (Scherer & Bänziger 2004).

Moreover, the use of prosody in sarcasm has not been studied systematically in different utterance types. Certain utterance types appear to be more readily used for sarcasm (e.g. positive declaratives) than other utterance types (e.g. negative declaratives, tag questions) (Kreuz & Glucksberg 1989, Kreuz & Caucci 2007). This raises the question of whether there may be a functional trade-off between prosodic and syntactic cues, as postulated in the Functional Hypothesis (Haan 2002). Stemming primarily from research on question intonation

in Dutch, this hypothesis argues that speakers use prosody to a lesser degree to convey a meaning in utterances that contain more syntactic and lexical cues to the same meaning than in utterances that contain fewer such cues. For example, *wh*-questions contain more lexical and syntactic markers of questions than yes-no questions and declarative questions and thus have fewer prosodic markers of questions that the latter in Dutch.

Finally, gender-related differences have been reported in the use of prosody to express affect and attitude (e.g. Van Leeuwen 1999). It remains to be investigated whether male and female speakers use prosody differently to express sarcasm.

To circumvent the limitations in previous analysis on sarcastic prosody, we adopt Functional Data Analysis (hereafter FDA; Ramsay & Silverman 2005), a new method for analysing continuous phonetic parameters, to model pitch variation in the key words in sarcastic utterances and their counterparts in sincere utterances as contours and represent the contours as continuous functions in statistical analysis (Gubian, Boves & Cangemi 2011, Jokisch, Langenberg & Pintér 2014, Gubian, Torreira & Boves 2015). Furthermore, to elucidate the use of prosody in different utterance types and gender-related differences, we take utterance type and gender into account in our experimental design.

2 The production experiment

2.1 Participants

Seventeen monolingual native speakers of the southern variety of British English (12 women, five men) participated in the production experiment. They were recruited from the student population of the University of Leeds, and between 18 and 23 years old (mean age = 21.6 years). They were paid a small fee for participation.

2.2 Tasks

We adapted the simulated telephone conversation task used in two studies of sarcastic prosody in second language learners of English (Chen & de Jong 2015, Smorenburg, Rodd & Chen 2015) to elicit sarcastic production in a dialogue setting. In this task, the participants were asked to imagine themselves on the phone chatting with a good friend. This was done to make the participants feel comfortable about using sarcastic speech, as previous studies have found that friends and familiar speakers are more likely to produce sarcasm (Gibbs 2000, Rockwell 2007, Bryant 2010). The friend, who was a native speaker of the southern variety of British English, made a number of remarks about fictional people and situations. The participants' task was to respond to the friend's remarks in a sarcastic manner. The utterances which they were supposed to say as the responses to the friend were displayed on PowerPoint slides (one slide per trial) to ensure consistency in choice of word and syntactic structure across participants.

Following the simulated telephone conversation task, the participants were asked to produce the response utterances in a sincere manner as responses to some imaginary remarks of the friend that would render a sincere manner appropriate. This time, only the response utterances were displayed on the PowerPoint slides and no context-setting remarks were played.

The participants were told in written instructions prior to the experiment that they could give an utterance a sarcastic or sincere sounding via prosody or the melody of speech. But they were not given definitions of 'sarcastic' and 'sincere'. The manner in which they were supposed to respond was displayed on the PowerPoint slides throughout the experiment. This was done because previous studies of verbal behaviour have shown that moderate emotions, such as interest, distress and sarcasm, are encoded more accurately when speakers are given explicit instructions to produce the intended emotion than otherwise (Rockwell 2000 and references therein).

2.3 Materials

The response utterances (to be produced by the participants) consisted of 48 utterances, representing three utterance types (16 utterances per type): simple (positive) declaratives (hereafter declaratives), utterances with negative question tags (hereafter tag questions), and *wh*-exclamatives. An example of each utterance type is given in (1)–(3), together with the context-setting remarks (only applicable in the 'sarcastic' condition).

(1) *Declaratives*
Remark: My aunt smokes a pack a day.
Response: She's a healthy lady.
(2) *Tag questions*
Remark: Kim turned up at my party even though she wasn't invited.
Response: You were pleased to see her, weren't you*?*
(3) Wh-*exclamatives*
Remark: My mother-in-law always smirks and snorts loudly when I misspeak.
Response: What a respectful gesture!

The context-setting remarks and response utterances (see the Appendix) were identical to the stimuli used in Smorenburg et al. (2015), which were partially derived from previous studies (Ackerman 1983, Kreuz & Glucksberg 1989, Capelli et al. 1990, Cheang & Pell. 2008, Chen & de Jong 2015). The response utterances were controlled for comparability in length and syntactic complexity. Three native speakers of British English, who had no connection to this study, were consulted to ensure equal acceptability of the response utterances being uttered sarcastically or sincerely on lexical and syntactic levels. The context-setting remarks were constructed carefully such that they created a convincing situation for the participants to respond in a sarcastic manner. However, in 19 of the cases, it was felt that the context-setting remarks alone might not sufficiently set the common ground between the participants and the English 'friend', which is supposed to facilitate the production of sarcasm (Kreuz et al. 1999). Additional written background information was thus provided in these cases.

The context-setting remarks in the 'sarcasm' condition were recorded by a male native speaker of the southern variety of British English. The speaker was provided with the context-setting remark–response sequences in the form of a recording script, and was asked to familiarize himself with the remarks. He was then asked to produce the context-setting remark-response sequences together with a female research assistant, a highly proficient learner of British English, who acted as the interlocutor and produced the responses. The male speaker was instructed to say the context-setting remarks as if he was having real conversations with the interlocutor and expecting responses from her on his remarks. The speakers produced the 48 context-setting remark–response sequences twice and recorded at a sampling frequency of 44.1 kHz (16 bits accuracy) in a sound-attenuated booth at the Linguistics Laboratory of the Utrecht Institute of Linguistics. During both rounds of recordings, the male speaker was encouraged to make additional attempts if he was not satisfied with his first attempt or if the first author, who was present during the recording, noticed inconsistency in the male speaker's prosody. The context-setting remarks from the second round of recording were selected and saved as individual .wav files using Praat (Boersma & Weenink 2014). In the case of multiple attempts, the last one was selected.

2.4 Procedure

The participants did the experiment individually in a sound-attenuated booth at the Phonetics Laboratory at the University of Leeds. Each participant was first presented with written instructions. In the instructions, the tasks and procedure of the experiment were explained. The participant was also informed that he/she was allowed to make multiple attempts on each

trial until he/she was satisfied with the response. The stimulus order was semi-randomised such that responses of the same utterance type did not occur twice in a row.

The experiment began with six practice trials (two per utterance type), comparable to the experimental trials, and proceeded first with the 48 trials in the sarcasm condition and then the 48 trials in the sincerity condition. It was self-paced and conducted using a laptop. Each context-setting remark (only applicable on the 'sarcastic' trials), the corresponding response, additional background information (only applicable on 19 of the 'sarcastic' trials) and the manner of response (sarcastic or sincere) were displayed in a PowerPoint slide on the laptop screen on each trial. The participants could hear the context-setting remark from the fictional friend via a headphone set by clicking on a sound icon placed next to the context-setting remark in the slide and respond to it accordingly. They could move on to the next trial by pressing the 'enter' key on the keyboard and took a short break during the experiment. Their responses were recorded using a ZOOM 1 digital recorder. It took on the average about 15 minutes for the participants to finish the experiment.

3 Data annotation and pre-processing

A total of 1529 utterances were obtained from 16 speakers (11 female, five male). The data of one female speaker was lost due to technical failure. Another seven utterances were judged unusable due to unexpected noise during the recording. To check whether the speakers produced the utterances as required in their respective condition, we conducted a small perception experiment. In this experiment, eight native speakers of English (four British English, three American English, one British English and American English bilingual) listened to 243 utterances from six speakers (three female, three male) and judged for each utterance whether the speaker intended to be sarcastic or sincere in a two-alternative forced choice task. We conducted mixed-effect binary logistic regression in SPSS (IBM SPSS version 22) to statistically assess which variables were predictive of the listeners' judgements (i.e. MESSAGE_TYPE_PERCEIVED). The predictors included three main effects, MESSAGE_TYPE_REQUIRED (sarcastic vs. sincere), i.e. the message type that the speakers were supposed to produce in the production experiment, UTTERANCE_TYPE (declarative, tag question, *wh*-exclamative) and VARIETY_OF_ENGLISH spoken by the listeners (British English vs. American English), three two-way interactions between these predictors and one three-way interaction of these predictors. The model could predict the intended message type of an utterance correctly in 73.3% of the sarcastic cases and 73.9% of the sincere cases. Only one of the predictors made a significant contribution to the model: MESSAGE_TYPE_REQUIRED ($\beta = -1.854$, $SE = 0.207$, $t = -8.964$, $p < .01$).

This result showed that the utterances were uttered by and large as they were supposed to in their respective conditions. Some utterances were perceived to sound opposite to what they were supposed to by some of the listeners. But the misperception of these utterances could be due to various reasons and itself did not suggest that the speakers did not produce the utterances according to the instructions in the production experiment. It is possible that the prosodic cues to sincerity and sarcasm might be weaker in these utterances than in other utterance. Such variation is, however, characteristic of prosody in natural speech. We therefore decided to include all the usable utterances into our analysis. These utterances were subsequently annotated and processed for further analysis in the following order.

First, the key word (see the underlined words in the Appendix) was identified for each sarcastic utterance in the light of the context-setting remarks and available background information. The key words in the sarcastic utterances and their counterparts in the sincere utterances were the words of interest (hereafter target words) in our prosodic and statistical analysis.

Second, each utterance was annotated for the begin and end of the target word by two annotators using Praat (Boersma & Weenink 2014). Each annotator annotated a subset of the speakers, two of which were annotated by both of them. An intra-class correlation coefficient (ICC) test was conducted on the annotation in the target words of the two speakers annotated by both annotators (N = 192). The ICC for the duration annotation was .993 ($F(191) = 142.451$, $p < .01$), indicating that the annotation done by the two annotator was highly consistent. The annotation of all the utterances was subsequently checked by one of the annotators and was adjusted when necessary.

Third, the tool ProsodyPro (Xu 2013) was used to extract from the target words the pitch and duration values at a sampling rate of 100 Hz (i.e. 10 ms spacing between successive points), including mean pitch, maximum pitch (pitch-max), minimum pitch (pitch-min), pitch span and word duration (duration). ProsodyPro exports continuous pitch values, even if the raw contour contain gaps due to the presence of voiceless sounds. Voiceless intervals are bridged by linear interpolation between the last pitch measurement before and the first pitch measurement after the gap.

Fourth, the pitch contours of the target words, which were of variable lengths and phonetic make-up, were normalised to have an equal number of samples, in preparation for FDA. The pitch contour of the shortest target word consisted of 12 samples of pitch values, while the pitch contour of the longest target word comprised 157 samples of pitch values. To create length-normalised contours, we re-sampled and smoothed the curves to obtain contours consisting of 50 samples each using spline interpolation (Gubian et al. 2015). Spline interpolation transforms sampled data representations to a continuous mathematical function that is guaranteed to be smooth and differentiable. For example, spline interpolation removes the sharp edges that may be caused by linear interpolation to close gaps related to voiceless intervals. Time normalisation by means of spline interpolation is assumed to capture the linguistically relevant aspects of the contour shapes, which are supposed to be independent of the length and phonetic make-up of the words.

Finally, we performed Functional Principal Component Analysis, a tool available in FDA, on the pitch contours of all target words, and extracted functional principal components (FPC) for each pitch contour for further analysis, following the procedure described in Gubian et al. (2015). Conventional Principal Component Analysis (PCA) identifies a small number of orthogonal dimensions on which the original observations can be represented with minimal loss of detail. Similarly, Functional PCA returns a number of continuous functions that can be used to reconstruct the original functions, with minimal loss of accuracy. In our case, it represents pitch contours by continuous spline functions. The FPCs are orthogonal, meaning that the integral of the product of any pair of FPCs is zero, reminiscent of Fourier decomposition. Individual observations are reconstructed by summing the mean function (common to all observations) and weighted versions of the FPCs. The result of the functional PCA analysis consists of the weights of the FPCs for each individual observation. This makes continuous time functions suitable for statistical analysis that operates on discrete numbers. At the same time, the weights represent the complete pitch contour, rather than a subset of hand-picked points on a pitch contour.

In our analysis, we decided to extract the first five FPCs for the pitch contour of each target word. These FPCs accounted for 94% of the total variance in the pitch contours of the target words (FPC1: 54.1%, FPC2: 23.6%, FPC3: 8.4%, FPC4: 4.8%, FPC5: 2.8% of the variance). We could have extracted additional FPCs, but these would have probably not been informative. It is already doubtful whether FPC5 is relevant. Figure 1 shows the shape of each FPC together with the average normalized pitch contour of the target words. To illustrate how the FPCs can modify a pitch contour, we show in Figure 2 how adding each FPC by a positive weight and a negative weight can change the shape of the average normalized pitch contour of the target words. More specifically, FPC1 was negative at the left end and positive at the right end of the pitch contour. A positive weight of FPC1 would lower the pitch at the left end of the average normalised pitch contour, and raise the pitch towards the right end; a negative weight of FPC1

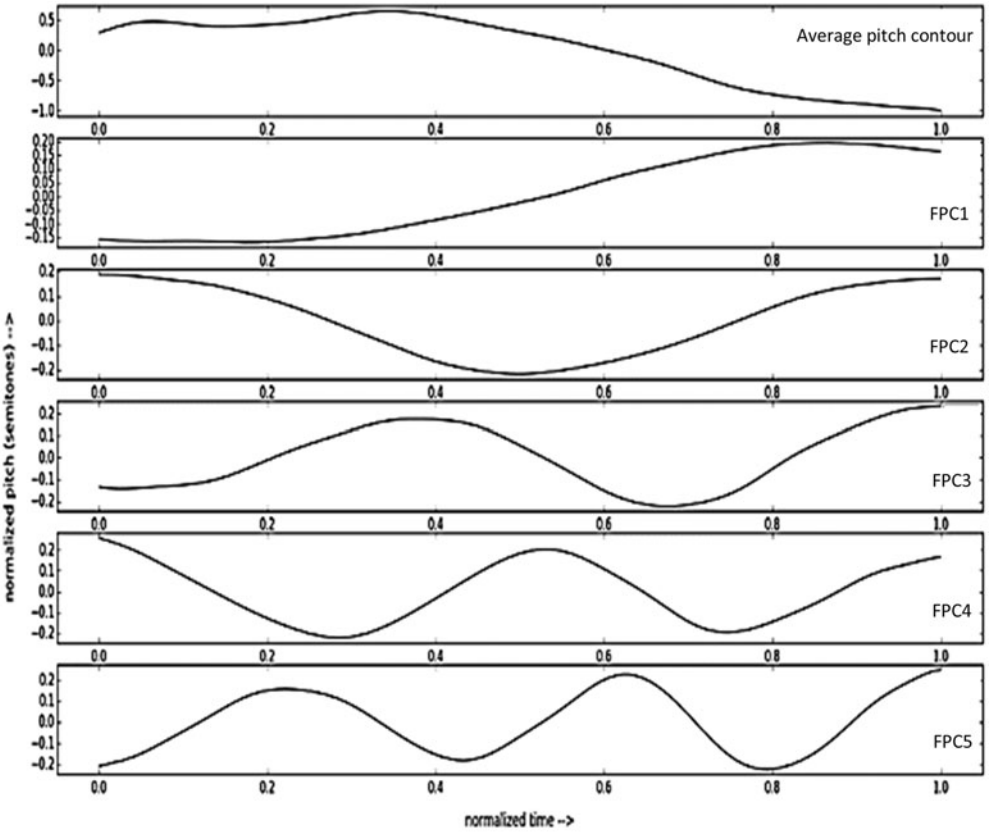

Figure 1 The average normalized pitch contour and the first five functional Principal Components of the target words (i.e. the key words in the sarcastic utterances and their counterparts in the sincere utterances). The horizontal axis corresponds to normalized time: $t = 0$ corresponds to the start of the target word, and $t = 1$ to the offset of the target word. The vertical axis corresponds to the normalized pitch in semitones, centered on zero.

would raise the pitch at the left end of the average normalized pitch contour and lower the pitch towards the right end. Since the average normalised contour was characterised by a fall in the second half of the contour, a positive weight of FPC1 would make the contour overall flatter and a negative weight of FPC1 would make the contour overall steeper (Figure 2). FPC2 was positive at the left end and the right end, and negative in the middle of the pitch contour. A positive weight of FPC2 would shift the onset of the pitch fall towards the left end of the contour and make the fall less steep; a negative weight of FPC2 would shift the onset of the fall towards the right end of the contour, and make the fall steeper (Figure 2). FPC3 strengthens the effect of FPC2 in alignment of the fall: a positive weight would result in an earlier fall, while a negative weight would result in a later fall. FPCs 4 and 5 add rapid pitch fluctuations to the average pitch contour. Their effect may have more to do with the perceived naturalness of the speech than with linguistic or paralinguistic information. It should be noted that the FPCs do not have an intrinsic scale. Their contribution to the reconstruction of a specific pitch contour is completely determined by the weight assigned by the Functional PCA to the FPCs for that contour.

The procedure described above resulted in ten measurements for statistical analysis: mean pitch, pitch-min, pitch-max, pitch span, duration, FPC1, FPC2, FPC3, FPC4, and FPC5.

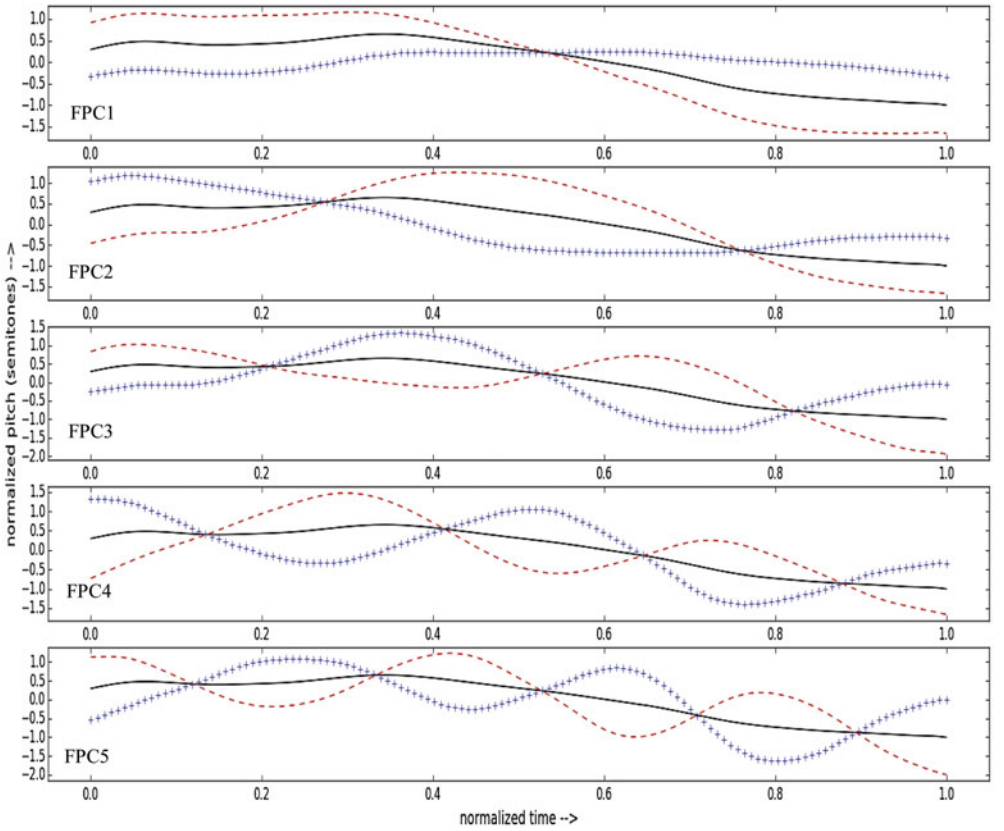

Figure 2 (Colour online) Modifications of the first five functional Principal Components to the normalized average pitch contour of the target words (i.e. the key words in the sarcastic utterances and their counterparts in the sincere utterances). In each panel, the solid line represents the normalized average pitch contour; the broken line and the plus-signed line represent the contours that would be obtained if the FPC would be added to the average with a weight of minus or plus 4 respectively. The horizontal axis corresponds to normalized time: $t = 0$ corresponds to the start of the target word, and $t = 1$ to the offset of the target word. The vertical axis corresponds to the normalized pitch in semitones, centered on zero.

4 Statistical analysis and results

4.1 Prosodic differences between sarcasm and sincerity

We first analysed effects of the experimental variables on each of the ten measurements (i.e. outcome variables) with linear mixed-effect models, using the *lme4* package (Bates & DebRoy 2006) in the R environment (R Development Core Team 2016). We included MESSAGE_TYPE (sarcastic vs. sincere), UTTERANCE_TYPE (declarative, tag question, *wh*-exclamative) and GENDER (male, female) as the fixed factors and SPEAKER and UTTERANCE as the random factors. For each outcome variable, six models were built, starting with a model with only the constant and random factors and adding one more term (i.e. a main effect or an interaction) to each subsequent model. The anova function in R was used to compare model fit of different models. The model with the best-fit was determined by comparing each model with the winning model in the preceding model comparisons. The summary of the best-fit model indicated which terms reached statistical significance. As we are primarily interested in the effect of MESSAGE_TYPE in interaction with the other fixed factors, we will only

report significant interactions between MESSAGE_TYPE and the other fixed factors, and in the absence of such interactions, significant main effect of MESSAGE_TYPE. In the case of interactions involving the factor MESSAGE_TYPE, we used mixed-effect models to analyse the effect of MESSAGE_TYPE in each utterance type or each gender to determine where the interactions came from.

4.1.1 Mean pitch

The best-fit model retained a significant interaction of MESSAGE_TYPE by GENDER ($p < .01$). Examining the effect of MESSAGE_TYPE in the production of male and female speakers separately, we found a significant main effect of MESSAGE_TYPE only in the female speakers' production ($\beta = 184.70$, $SE = 16.13$, $t = 11.45$, $p < .01$). The female speakers used a significantly lower mean pitch in the sarcastic condition (201 Hz) than in the sincere condition (229 Hz), whereas the male speakers used similar pitch means in these conditions (124 Hz in the sarcastic condition vs. 131 Hz in the sincere condition). As can be seen in Figure 3, the pitch contours of the target words (*biting*, *smart* and *brilliant*) were positioned clearly lower in the pitch plot in the sarcastic condition than in the sincere condition in the female speakers (middle and bottom panels) but this was not the case in the male speaker (top panel).

4.1.2 Pitch-min

The best-fit model retained a significant interaction of MESSAGE_TYPE by UTTERANCE_TYPE ($p < .01$). Subsequent analyses revealed a significant main effect of MESSAGE_TYPE in each utterance type ($\beta = 12.67$, $SE = 3.19$, $t = 3.97$, $p < .01$ in declaratives; $\beta = 14.68$, $SE = 3.05$, $t = 4.85$, $p < .01$ in tag questions; $\beta = 29.29$, $SE = 4.16$, $t = 7.05$, $p < .01$ in *wh*-exclamatives). The speakers used a significantly lower pitch-min in the sarcastic condition than in the sincere condition across utterance types. Importantly, the difference in pitch-min was more pronounced in the *wh*-exclamatives (144 Hz in the sarcastic condition vs. 173 Hz in the sincere condition) than in the tag questions (133 Hz in the sarcastic condition vs. 148 Hz in the sincere condition) and declaratives (126 Hz in the sarcastic condition vs. 136 Hz in the sincere condition), as illustrated by examples in Figure 3.

4.1.3 Pitch-max

The best-fit model retained a significant interaction of MESSAGE_TYPE by UTTERANCE_TYPE ($p < .05$). Subsequent analyses revealed a significant main effect of MESSAGE_TYPE in the tag questions ($\beta = 28.21$, $SE = 4.51$, $t = 6.26$, $p < .01$) and *wh*-exclamatives ($\beta = 24.53$, $SE = 5.68$, $t = 4.32$, $p < .01$). The speakers used a significantly lower pitch-max in the sarcastic condition than in the sincere condition in these two utterance types (203 Hz in the sarcastic condition vs. 231 Hz in the sincere condition in the tag questions; 239 Hz in the sarcastic condition vs. 262 Hz in the sincere condition in the *wh*-exclamatives) but not in the declaratives (208 Hz in the sarcastic condition vs. 216 Hz in the sincere condition), as illustrated by examples in Figure 3.

4.1.4 Pitch span

The best-fit model did not retain any significant main effects and interactions for any of the fixed factors for variation in pitch span. The speakers thus did not vary pitch span systematically to express sarcasm.

4.1.5 Duration

The best-fit model retained a significant interaction of MESSAGE_TYPE by GENDER ($p < .01$). Subsequent analyses revealed a significant main effect of MESSAGE_TYPE in both the male ($\beta = -97.23$, $SE = 8.85$, $t = -10.99$, $p < .01$) and female speakers ($\beta = -58.68$, $SE = 5.51$, $t = -10.65$, $p < .01$). Both groups of speakers used a significantly longer duration in the sarcastic condition than in the sincere condition, but the male speakers exhibited a larger duration difference (507 ms in the sarcastic condition vs. 407 ms in the sincere condition) than the

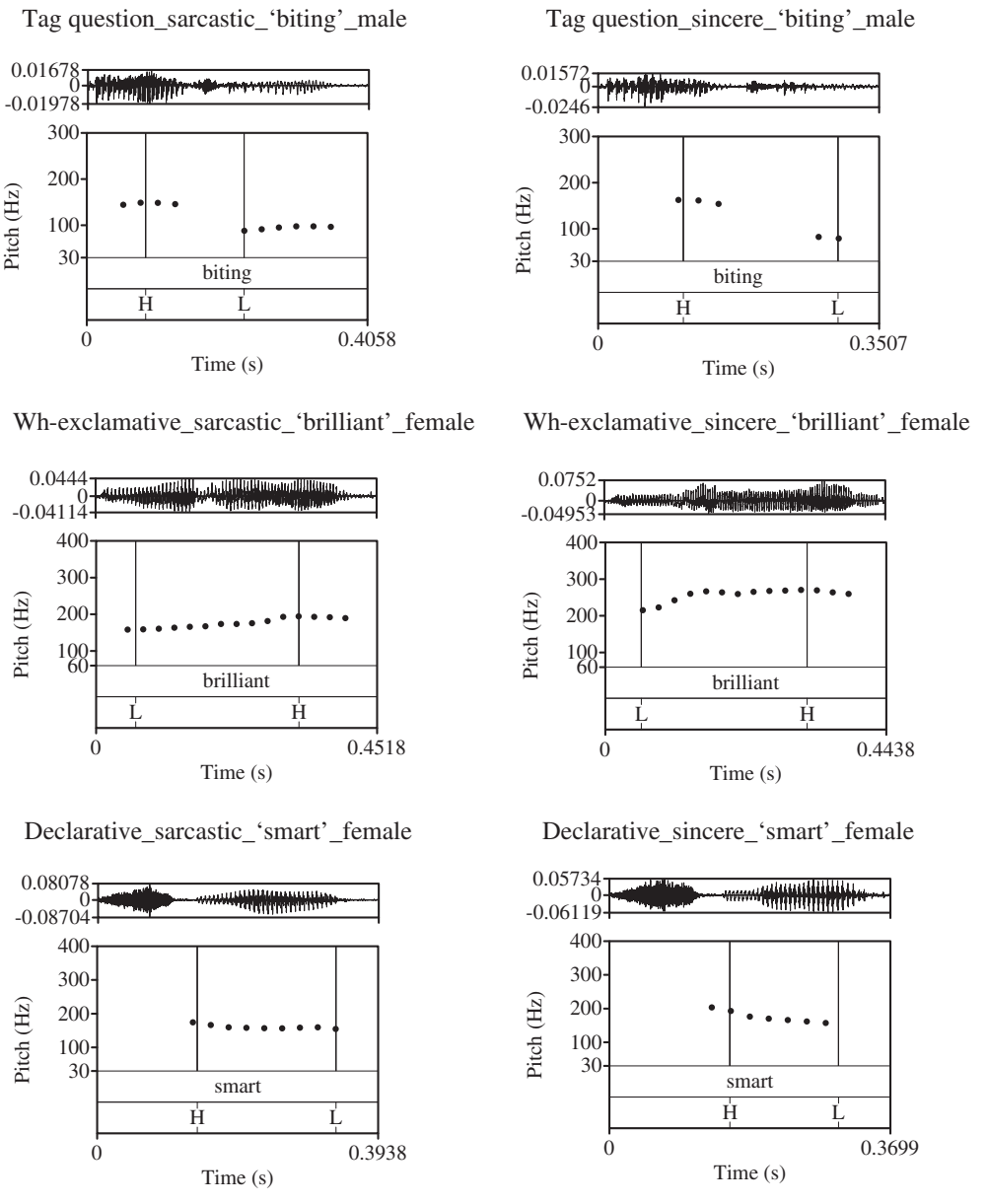

Figure 3 Examples of the pitch contours of the target words produced in three utterance types (top panels: tag question, middle panels: *wh*-exclamative; bottom panels: declarative) in the sarcastic (left panels) and sincere (right panels) conditions by one male speaker (L34) (top panel) and one female speaker (L05) (middle and lower panels). The target word *biting* was from the utterance 'They're biting this season, aren't' they?'. The target word *brilliant* was from the utterance 'What a brilliant header!'. The target word *smart* was from the utterance 'You're a smart gamer'. 'H' stands for pitch-max, 'L' stands for pitch-min'. The horizontal axis shows the duration of each word in seconds; the vertical axis shows the amplitude in Pascals in the waveform and pitch in Hz in the pitch contour.

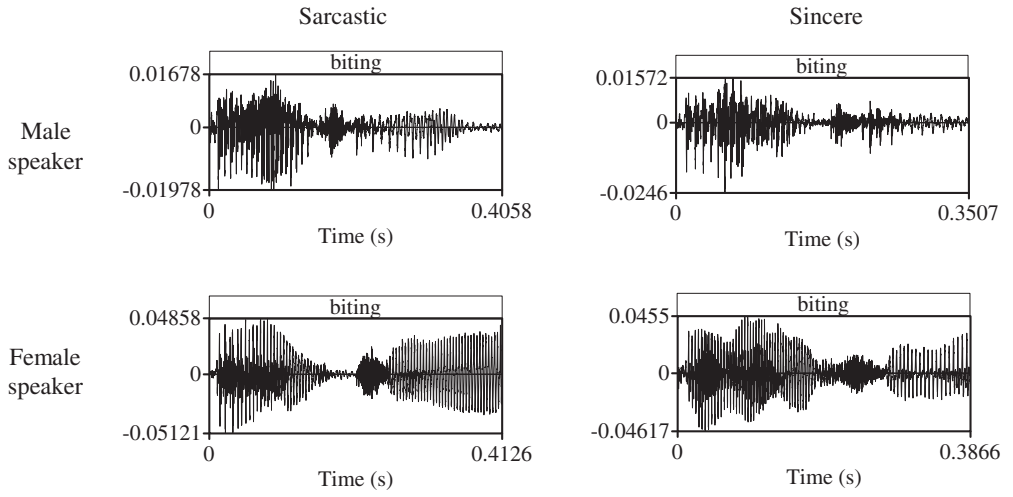

Figure 4 Renditions of the target word *biting* in the utterance 'They're biting this season, aren't they?' produced by a male speaker (L34) (top panels) and a female speaker (L03) (bottom panels) in the 'sarcastic' condition (left panels) and the 'sincere' condition' (right panels). The duration difference in *biting* between the two conditions was 56 ms in the male speaker but 26 ms in the female speaker. The horizontal axis shows the duration of each word in seconds; the vertical axis shows the amplitude in Pascals.

female speakers (521 ms in the sarcastic condition vs. 463 ms in the sincere condition), as illustrated in the examples in Figure 4.

4.1.6 FPC1

The best-fit model retained a significant main effect of MESSAGE_TYPE ($\beta = -4.85$, $SE = 1.26$, $t = -3.84$, $p < .01$). FPC1 was positive (2.19) in the sarcastic condition but negative in the sincere condition (–2.19). That is, the pitch contours of the target words had a flatter fall in the sarcastic condition than in the sincere condition.

4.1.7 FPC3

The best-fit model retained a significant interaction of MESSAGE_TYPE by UTTERANCE_TYPE ($p < .01$). Subsequent analyses revealed a significant main effect of MESSAGE_TYPE only in the tag questions ($\beta = 1.42$, $SE = 0.48$, $t = 2.93$, $p < .01$). FPC3 was negative in the sarcastic condition (–0.34) and positive in the sincere condition (1.09) in these utterances. That is, the contours of the target words in the tag questions had a later fall in the sarcastic condition than in the sincere condition

4.1.8 FPCs 2, 4 and 5

The best-fit model did not retain any significant main effects and interactions for any of the fixed factors regarding FPCs 2, 4 and 5.

4.1.9 Interim summary

Our linear mixed-effect models show that six of the ten measurements (mean pitch, pitch-min, pitch-max, duration, FPC1, FPC3) were significantly different between the sarcastic condition and the sincere condition. However, the measurements mean pitch and duration were not varied to the same degree in both groups of speakers. Specifically, only the female speakers used a significantly lower mean pitch in the sarcastic condition than in the sincere condition; the male speakers produced a significantly larger difference in duration between the sarcastic and sincere conditions than the female speakers. Further, the measurements pitch-min, pitch-max and FPC3 were not varied to the same degree in all utterance types.

Table 1 Summary of the results of the mixed-effect binary logistic regression model on the prediction that an utterance was uttered with sarcasm. The reference category was 'not sarcastic' for the outcome variable, male for the fixed factor gender, and *wh*-exclamative for the fixed factor utterance_type.

				95% C.I. for EXP(B)	
	Coefficient estimate	Std.Error	Sig.	Lower	Upper
Intercept	−4.821	1.868	.100	−8.485	−1.156
Gender (female)	4.190	2.130	.049	0.012	8.367
FPC3	0.118	0.053	.027	0.013	0.223
FPC5	0.170	0.002	< .001	0.010	0.331
Duration	0.010	0.002	< .001	0.007	0.014
FPC3 * gender					
FPC3 by gender (female)	−0.122	0.058	.034	−0.235	−0.009
Duration * gender					
Duration by gender (female)	−0.004	0.002	.043	−0.007	−0.0001
Duration * utterance_type					
Duration by utterance_type (decl)	−0.005	0.002	.020	−0.009	−0.001
Duration by utterance_type (tag)	0.0001	0.003	.963	−0.005	0.006

Specifically, the speakers lowered pitch-min in the sarcastic condition to a larger degree in the *wh*-exclamatives than in tag questions and declaratives, used a significantly lower pitch-max in the sarcastic condition than in the sincere condition only in *wh*-exclamatives and tag questions, and a later fall in the sarcastic condition than in the sincere condition only in the tag questions.

4.2 Predicting sarcasm vs. sincerity

Although most of the measurements have turned out to differ between the sarcastic condition and sincere condition, this does not necessarily mean that these measurements would be similarly useful to predict whether an utterance was uttered with sarcasm or not. We assessed the predictive power of these measurements using the mixed-effect binary logistic regression model in SPSS (IBM SPSS version 22). The outcome variable of the model was MESSAGE_TYPE (sarcastic vs. not sarcastic). The predictor variables included 12 main effects (i.e. GENDER, UTTERANCE_TYPE, MEAN PITCH, PITCH-MIN, PITCH-MAX, PITCH SPAN, DURATION, FPC1, FPC2, FPC3, FPC4 and FPC5), 21 two-way interactions (i.e. GENDER by UTTERANCE_TYPE, and the interaction of each continuous predictor by each categorical predictor), and ten three-way interactions between the two categorical predictors (GENDER, UTTERANCE_TYPE) and each continuous predictor. In addition, two random variables were added to the model, i.e. SPEAKER and UTTERANCE.

The model could predict the message type of an utterance correctly in 73.7% of the cases. Seven of the predictor variables made a significant contribution to the model: GENDER ($p < .05$), FPC3 ($p < .05$), FPC5 ($p < .05$), DURATION ($p < .01$), FPC3 by GENDER ($p < .05$), DURATION by GENDER ($p < .05$) and DURATION by UTTERANCE_TYPE ($p < .05$). This result shows that not all the measurements in which sarcastic and sincere conditions differ are equally useful in predicting the presence of sarcasm. Since the predictors GENDER, FPC3 and DURATION were involved in the interactions, we focus on the effects of FPC5 and the three interactions here (Seltman 2015). First, as shown by the coefficient estimates (Table 1), every unit of increase in FPC5 led to an increase by 0.17 in the odds of an utterance being interpreted to sound sarcastic (hereafter the odds of sarcasm). Second, although a unit of increase in FPC3 led to an increase in the odds of sarcasm in general (main effect of FPC3), this increase was smaller for the female speakers than for the male speakers by 0.122. Similarly, although a unit of increase in duration led to an increase in the odds of sarcasm in general (main effect

of DURATION), the increase the odds of sarcasm was smaller in the declaratives than in the tag questions and *wh*-exclamatives.

5 Discussion and conclusions

Taking a linguistically-motivated approach, we have examined prosodic expression of sarcasm in the key word of an utterance (i.e. the word that is semantically critical word to the expression of sarcasm) with respect to both static and dynamic prosodic measurements, and the predictive power of prosodic and non-prosodic factors for the presence (or absence) of sarcasm in the southern variety of British English. We have found that the target words (i.e. the key words in sarcastic utterances and their counterparts in the sincere utterances) are realised with a longer duration and a flatter pitch contour in the sarcastic condition than in the sincere condition regardless of utterance type and speaker gender. The longer duration of the target words in sarcastic utterances may suggest a slower speaking rate at the utterance level, as reported for American and Canadian English (Rockwell 2000, Cheang & Pell 2008), Cantonese (Cheang & Pell 2009), Mexican Spanish (Rao 2013) and French (Loevenbruck et al. 2013, González-Fuente et al. 2016). The flatter pitch contour appears to be compatible with a smaller pitch span at the utterance level in sarcastic utterance, as reported for Cantonese (Cheang & Pell 2009), American English (Rakov & Rosenberg 2013), Mexican Spanish (Rao 2013) and German (Niebuhr 2014), but different from the reported use of a larger pitch span in Italian (Anolli et al. 2002) and French (González-Fuente et al. 2016, see also Loevenbruck et al. 2013) and more dynamic pitch contours in Catalan (González-Fuente et al. 2015). Sarcastic speech in British English thus appears to be largely similar to North American English and differs from other languages in pitch span.

More importantly, our study has yielded insights into aspects of prosodic expression of sarcasm that have not been studied in previous work on English and other languages. First, we have found notable differences in the use of prosody in different utterance types. Specifically, the target words are realised with a lower pitch-min in the sarcastic condition than in the sincere condition across utterance types but the difference in pitch-min is more pronounced in *wh*-exclamatives than in tag questions and declaratives. Further, they are realised with a lower pitch-max in the sarcastic condition than in the sincere condition in *wh*-exclamatives and tag questions but not in declaratives. The target words are also realised with a later fall in the sarcastic condition than in the sincere condition in tag questions. Taking account of the findings on duration and flatness of pitch contour mentioned in the preceding paragraph, we may suggest that sarcasm and sincerity differ in five of the prosodic parameters under examination in tag questions (i.e. duration, pitch-min, pitch-max, FPC1, FPC3), in four of the prosodic parameters in *wh*-exclamatives (i.e. duration, pitch-min, pitch-max, FPC1), and three of the prosodic parameters in declaratives (i.e. duration, pitch-min, FPC1). It has been claimed that positive declaratives may be more readily used sarcastically than tag questions (Kreuz & Glucksberg 1989, Kreuz & Caucci 2007). Our results may thus suggest a functional trade-off between the readiness of an utterance type being used sarcastically and the presence of prosodic cues to sarcasm, in line with the Functional Hypothesis (Haan 2002). However, as independent empirical evidence for differences in how readily different utterance types can be used sarcastically is still lacking, future research is needed to validate our interpretation of the results on the presence of prosodic cues in different utterance types.

Interestingly, we have also observed gender-related differences in prosodic expression of sarcasm. Specifically, the target words in the sarcastic condition are realised with a lower mean pitch than in the sincere condition by female speakers, not by male speakers. This finding raises the question as to whether the reported use of a lower mean pitch in sarcastic utterances at the utterance level in North American English is primarily applicable to female speakers. The target words are also realised with a longer duration in the sarcastic condition

than in the sincere condition by both male and female speakers but the duration difference between the two conditions is much larger in male speakers' production. The more extensive use of duration in male speakers may serve as a compensation for a lack of use of mean pitch.

Additionally, we have found that the prosody of the target words has yielded useful predictors for the presence (or absence) or sarcasm, i.e. FPC3, FPC5 and DURATION. Together with the non-prosodic factors UTTERANCE_TYPE and GENDER, they can accurately predict the presence (or absence) of sarcasm in nearly 74% of the time. To get an idea of the predictive power of the prosodic measurements alone, we conducted the mixed-effect binary logistic regression analysis excluding the non-prosodic predictors UTTERANCE_TYPE and GENDER. The new model (containing MEAN PITCH, PITCH-MIN, PITCH-MAX, PITCH SPAN, DURATION, FPC1, FPC2, FPC3, FPC4 and FPC5) could make correct predictions in 70.4% of the cases, suggesting that the prosodic predictors are primarily responsible for the improved performance of the best-fit model, compared to the intercept-only model.

Together, our results show that sarcasm and sincerity can be prosodically distinguishable in the key words alone and the presence of sarcasm can be predicted with a reasonable accuracy with prosodic information from the key words in British English, lending support to our key-word–based approach. Arguably, the key words may also be analysed as the focus in the sarcastic utterances (narrow focus). Its counterpart in the sincere utterances may be considered part of the whole-utterance focus (or broad focus), as the sincere utterances were produced without explicit contexts. It can thus not be ruled out that the prosodic differences found between the sarcastic and sincere conditions might in part be attributed to the differences between narrow focus and broad focus at least on some trials and in some speakers. Future studies in which focus conditions of sarcasm and sincere utterances are systematically varied are needed to shed light on the interface between sarcasm and focus in prosody.

Our results also imply that the key words do not contain all the prosodic variation relevant to the expression of sarcasm, and that not all relevant prosodic variation is encoded in conventional prosodic parameters. Informal listening to the utterances suggested that sarcasm may also be signalled by subtle changes in the overall pitch contour of the complete utterance in British English. In follow-up research we will investigate this issue more formally. However, it is not straightforward to apply functional PCA to complete utterances in the current data set, mainly because of the presence of variable numbers and lengths of voiceless intervals. While we could apply functional PCA to the target words, despite the fact that many of the target words contained one or more unvoiced intervals, we will explore whether more advanced interpolation methods for handling voiceless intervals can improve the results of FDA, for example, by taking the phonetic make-up of the words and utterances into account. Informal listening also suggested that voice quality may play a role in the expression of sarcasm in British English, as found for German (Niebhur 2014). Many voice quality features (Laver 1980, Boves 1984, Van Bezooijen 1984) can be expressed as continuous functions, and made accessible to functional PCA processing. In future research, we will extend the use of FDA from pitch to voice quality and examine the use of different voice quality features at both the key-word level and the utterance level in the expression of sarcasm in British English.

Acknowledgements

We are grateful to Laura Smorenburg and Joe Rodd for their assistance in recording the context-setting remarks used in the production experiment, and preparing and administering the production experiment in the UK, Laura Smorenburg for preparing and administering the perception experiment, Cécile de Cat for allowing us to use the Phonetics Lab at the University of Leeds to conduct the production experiment, Karlijn Blommers and Rianne Kamerbeek for data annotation, and Karlijn Blommers for editing the references. We also wish to thank Santiago González-Fuente and four anonymous reviewers for their thorough reviews and constructive feedback. This research was conducted with the support of an Aspasia grant (grant number 015.007.013) awarded to the first author by the Netherlands Organisation for Scientific Research (NWO).

Appendix. Context-setting remarks and response utterances

The context-setting remarks and the response utterances in the sarcastic condition with the semantically crucial word or the key word underlined (Decl: declarative, Tag: tag question, Wh: *wh*-exclamative

Trial	Remark from friend	Response from participant	Utterance_ type
1	My aunt smokes a pack a day.	She's a healthy lady.	Decl
2	I went for a run and I came back dripping wet.	It's a beautiful day outside.	Decl
3	I haven't seen a waiter yet since we were shown a table half an hour ago.	The service's really good here.	Decl
4	My plane was an hour late.	Ryanair's always reliable.	Decl
5	I bought a new game and I'd thought it would be too hard, but I learned in five minutes.	You're a smart gamer.	Decl
6	Tomorrow's class is going to be about plants.	That'll be great fun.	Decl
7	I heard Peter, that skinny kid with glasses, is gonna beat you up after school.	That's very scary.	Decl
8	My brother was accepted to the police academy.	Your parents must be proud.	Decl
9	I got one of the lowest grades in the maths test.	That's never happened before.	Decl
10	I heard that the camping trip has been cancelled.	You must be really disappointed.	Decl
11	My uncle keeps telling that stupid joke over and over again.	That joke's hilarious.	Decl
12	My baby sister fell asleep on her dinner plate.	That sounds comfortable.	Decl
13	My dance partner keeps stepping on my toes.	She's a graceful dance partner.	Decl
14	Only three people showed up to my housemate's party last night.	That sounds wild.	Decl
15	How's being stuck at home after your accident?	I'm having a great time.	Decl
16	Did you see my art coursework?	You could be a professional artist.	Decl
17	I traded my cricket bat for a toy truck, but now I find out it's broken.	You got a bargain, didn't you?	Tag
18	My brother wanted to help me move and he dropped my grandfather's clock.	He was a big help, wasn't he?	Tag
19	I put my homework off for two hours, but then it only took ten minutes.	That took a lot of effort, didn't it?	Tag
20	I think I didn't even get one right on that test.	You did well this time, didn't you?	Tag
21	My little sister kicked me in the shins.	Your sister's sweet, isn't she?	Tag
22	My first football training coach made us run 5 miles; some of the guys threw up.	You lot are in shape, aren't you?	Tag
23	I went fishing but didn't catch anything.	They're biting this season, aren't they?	Tag
24	I've joined a running club and go for a run every evening.	You're a devoted athlete, aren't you?	Tag
25	I warned Mark about washing his whites and coloureds together.	He takes advice well, doesn't he?	Tag
26	The weather forecast didn't say it would rain today.	The weather forecast is always right, isn't it?	Tag
27	My nephew showed his brand new iPhone to everyone at the party.	He's modest, isn't he?	Tag
28	I had a busy morning, walking the dog and doing the dishes	You've worked hard, haven't you?	Tag
29	Kim turned up at my party even though she wasn't invited.	You were pleased to see her, weren't you?	Tag
30	My father helped me paint my flat; there's paint splatters everywhere.	He's done a beautiful job, hasn't he?	Tag
31	I had to serve all my aunts and uncles drinks all afternoon.	They were gracious guests, weren't they?	Tag
32	I caught my cat eating the fish from next door's pond.	You're a nice neighbour, aren't you?	Tag
33	My mother-in-law always smirks and snorts loudly when I misspeak	What a respectful gesture!	Wh
34	The arrogant front-runner finished dead last.	What an amazing result!	Wh
35	Today, playing football, I slipped and fell and the ball bounced off my head.	What a brilliant header!	Wh
36	My piano performance has been cancelled.	What a terrible shame!	Wh
37	My sister phoned just now and told me that her job interview went very badly.	What a surprising outcome!	Wh
38	My father had made a five course Christmas dinner; the next day we were all sick with food poisoning.	What an accomplished chef!	Wh
39	I failed my essay on parliamentary process. I guess I just don't know much about politics.	What a shocking announcement!	Wh
40	My grandma bought me a yellow sweater for my birthday.	What a lovely colour!	Wh

Trial	Remark from friend	Response from participant	Utterance_type
41	My cat left me a present on the doormat.	What a nice surprise!	Wh
42	I ordered soup for lunch and found a hair in it.	What a tasty lunch!	Wh
43	I can touch my nose with my elbow.	What a useful skill!	Wh
44	My sister's boyfriend ran out of the haunted house screaming like a little girl.	What a brave man!	Wh
45	I got a lift from Mary. She kept indicating the wrong way at roundabouts.	What a great driver!	Wh
46	People left the cinema halfway through the film.	What a gripping film!	Wh
47	My brother's friend came to our house just because he wanted to play our new game.	What a considerate friend!	Wh
48	I almost fell asleep in class today.	What an engaging lecture!	Wh

References

Ackerman, Brian P. 1983. Form and function in children's understanding of ironic utterances. *Journal of Experimental Child Psychology* 35(3), 487–508.

Anolli, Luigi, Rita Ciceri & Maria G. Infantino. 2000. Irony as a game of implicitness: Acoustic profiles of ironic communication. *Journal of Psycholinguistic Research* 29(3), 275–311.

Anolli, Luigi, Rita Ciceri & Maria G. Infantino. 2002. From "blame by praise" to "praise by blame": Analysis of vocal patterns in ironic communication. *International Journal of Psychology* 37, 266–276.

Attardo, Salvatore, Jodi Eisterhold, Jennifer Hay & Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor: International Journal of Humor Research* 16(2), 243–260.

Bates, Douglas & Sarkar DebRoy. 2006. *lme4*: Linear mixed-effects models using S4 classes. http://CRAN.R-project.org, R package version 0.99875-8 (accessed 1 May 2016).

Boersma, Paul & David Weenink. 2014. Praat: Doing phonetics by computer [computer program]. Version 5.3.62. http://www.praat.org/ (accessed 1 April 2016).

Boves, Lou. 1984. *The phonetic basis of perceptual ratings of running speech*. Dordrecht: Foris.

Brown, Penelope & Stephen C. Levinson. 1987. *Politeness: Some universals in language use*. Cambridge & New York: Cambridge University Press.

Bryant, Gregory A. 2010. Prosodic contrasts in ironic speech. *Discourse Processes* 47(7), 545–566.

Capelli, Carol A., Noreen Nakagawa & Cary M. Madden. 1990. How children understand sarcasm: The role of context and intonation. *Child Development* 61(6), 1824–1841.

Cheang, Henry S. & Marc D. Pell. 2008. The sound of sarcasm. *Speech Communication* 50(5), 366–381.

Cheang, Henry S. & Marc D. Pell. 2009. Acoustic markers of sarcasm in Cantonese and English. *The Journal of the Acoustical Society of America* 126(3), 1394–1405.

Chen, Aoju. 2012. Prosodic investigation on information structure. In Manfred Krifka & Renate Musan (eds.), *The expression of information structure*, 251–286. Berlin & New York: Mouton de Gruyter.

Chen, Aoju & Diantha de Jong. 2015. Prosodic expression of sarcasm in L2 English. In Marina Chini (ed.), *L2 spoken discourse: Pragmatic and prosodic aspects*, 27–37. Milano: FrancoAngeli.

Clark, Herbert H. & Richard J. Gerrig. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General* 113(1), 121–126.

Clift, Rebecca. 1999. Irony in conversation. *Language in Society* 28(4), 523–553.

Dumouchel, Pierre, Najim Dehak, Yazid Attabi, Reda Dehak & Narjès Boufaden. 2009. Cepstral and long-term features for emotion recognition. *Proceedings of the 10th Annual Conference of the International Speech Communication Association* (INTERSPEECH 2009), Brighton, 344–347.

Gibbs, Raymond W. 1999. Speakers' intuitions and pragmatic theory. *Cognition* 69(3), 355–359.

Gibbs, Raymond W. 2000. Irony in talk among friends. *Metaphor and Symbol* 15(1), 5–27.

González-Fuente, Santiago, Victoria Escandell-Vidal & Pilar Prieto. 2015. Gestural codas pave the way to the understanding of verbal irony. *Journal of Pragmatics* 90, 26–47.

González-Fuente, Santiago, Pilar Prieto & Ira Noveck. 2016. A fine-grained analysis of the acoustic cues involved in verbal irony recognition in French. *Proceedings of the 8th International Conference on Speech Prosody* (SP2016), Boston, MA, University of Boston, 902–906.

Grabe, Esther, Greg Kochanski & John Coleman. 2007. Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech* 50(3), 281–310.

Gubian, Michele, Lou Boves & Francesco Cangemi. 2011. Joint analysis of F0 and speech rate with functional data analysis. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 2011), Prague Congress Center, Prague, vol. 2, 1–4.

Gubian, Michele, Francisco Torreira & Lou Boves. 2015. Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics* 49, 16–40.

Haan, Judith J. M. 2002. *Speaking of questions: An exploration of Dutch question intonation*. Utrecht: LOT.

Jokisch, Oliver, Tristan Langenberg & Gábor Pintér. 2014. Intonation-based classification of language proficiency using FDA. *Proceedings of the 7th International Conference on Speech Prosody* (SP2016), Dublin, Trinity College Dublin, 795–798.

Jorgensen, Julia. 1996. The functions of sarcastic irony in speech. *Journal of Pragmatics* 26(5), 613–634.

Knox, Norman. 1961. *The word irony and its context: 1500–1755*. Durham, NC: Duke University Press.

Kreuz, Roger J. 2000. The production and processing of verbal irony. *Metaphor and Symbol* 15(1), 99–107.

Kreuz, Roger J. & Gina M. Caucci. 2007. Lexical influences on the perception of sarcasm. *Proceedings of the Workshop on Computational Approaches to Figurative Language*, 1–4.

Kreuz, Roger J. & Sam Glucksberg. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General* 118(4), 374–386.

Kreuz, Roger J., Max A. Kassler, Lori Coppenrath & Bonnie M. Allen. 1999. Tag questions and common ground effects in the perception of verbal Irony. *Journal of Pragmatics* 31, 1685–1700.

Ladd, D. Robert. 1996. *Intonational phonology*. Cambridge & New York: Cambridge University Press.

Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge & New York: Cambridge University Press.

Laver, John. 1980. *The phonetic description of voice quality*. Cambridge & New York: Cambridge University Press.

Lee, Christopher J. & Albert N. Katz. 1998. The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol* 13(1), 1–15.

Loevenbruck, Hélène, Mohamed A. Ben Jannet, Mariapaola D'Imperio, Mathilde Spini & Maud Champagne-Lavau. 2013. Prosodic cues of sarcastic speech in French: Slower, higher, wider. *Proceedings of the 14th Annual Conference of the International Speech Communication Association* (INTERSPEECH 2013), Lyon, 3537–3541.

Muecke, Douglas C. 1980. *The compass of irony*. London: Methuen.

Niebuhr, Oliver. 2014. "A little more ironic": Voice quality and segmental reduction differences between sarcastic and neutral utterances. *Proceedings of the 7th International Conference on Speech Prosody* (SP2016), Dublin, Trinity College Dublin, 608–612.

Oring, Elliott. 1994. Humor and the suppression of sentiment. *Humor* 7(1), 7–26.

Pierrehumbert, Janet B. 1980. *The phonetics and phonology of English intonation*. Ph.D. thesis, MIT.

R Development Core Team. 2016. R: A language and environment for statistical computing, reference index (version 3.2.5). Vienna: R Foundation for Statistical Computing. http://www.R-project.org (accessed 1 May 2016).

Rakov, Rachel & Andrew Rosenberg. 2013. 'Sure, I did the right thing': A system for sarcasm detection in speech. *Proceedings of the 14th Annual Conference of the International Speech Communication Association* (INTERSPEECH 2013), Lyon, 842–846.

Ramsay, J. O. & B. W. Silverman. 2005. *Functional data analysis*, 2nd edn. New York: Springer.

Rao, Rajiv. 2013. Prosodic consequences of sarcasm versus sincerity. *Concentric: Studies in Linguistics* 2, 33–59.

Rockwell, Patricia. 2000. Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic Research* 29(5), 483–495.

Rockwell, Patricia. 2001. Facial expression and sarcasm. *Perceptual and Motor Skills* 93(1), 47–50.

Rockwell, Patricia. 2007. Vocal features of conversational sarcasm: A comparison of methods. *Journal of Psycholinguistic Research* 36(5), 361–369.

Scherer, Klaus R. & Tanja Bänziger. 2004. Emotional expression in prosody: A review and an agenda for future research. *Proceedings of the 2nd International Conference on Speech Prosody* (SP2004), Nara, 359–366.

Seltman, Howard J. 2015. *Experimental design and analysis*. http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf (accessed 24 November 2017).

Smorenburg, Laura, Joe Rodd & Aoju Chen. 2015. The effect of explicit training on the prosodic production of L2 sarcasm by Dutch learners of English. *Proceedings of the 18th International Congress of Phonetic Sciences* (ICPhS 2015), Glasgow, 1–47.

Sperber, Dan. 1984. Verbal irony: Pretense or echoic mention? *Journal of Experimental Psychology: General* 113(1), 130–136.

Tepperman, Joseph, David Traum & Shrikanth Narayanan. 2006. "Yeah right": Sarcasm recognition for spoken dialogue systems. *Proceedings of the 9th Annual Conference of the International Speech Communication Association* (INTERSPEECH 2006), Pittsburgh, PA, 1838–1841.

Vallduví, Enric & Elisabet Engdahl. 1996. The linguistic realization of information packaging. *Linguistics* 34(3), 459–519.

Van Bezooijen, Renée.1984. *Characteristics and recognizability of vocal expressions of emotion*. Berlin & Boston, MA: Mouton De Gruyter.

Van Leeuwen, Theo. 1999. *Speech, music, sound*. London: Palgrave Macmillan.

Ward, Nigel G. & David DeVault. 2017. Challenges in building highly-interactive dialog systems. *AI Magazine* 37(4), 7–18.

Xu, Yi. 2013. ProsodyPro – A tool for large-scale systematic prosody analysis. *Proceedings of Tools and Resources for the Analysis of Speech Prosody* (TRASP 2013), Aix-en-Provence, 7–10.