When “okay” is not okay: Acoustic characteristics of single-word prosody conveying reluctance

Marianne van Zyl, Johan J. Hanekom, and AL

Citation: The Journal of the Acoustical Society of America **133**, EL13 (2013); doi: 10.1121/1.4769399
View online: http://dx.doi.org/10.1121/1.4769399
View Table of Contents: http://asa.scitation.org/toc/jas/133/1
Published by the Acoustical Society of America

When "okay" is not okay: Acoustic characteristics of single-word prosody conveying reluctance

Marianne van Zyl and Johan J. Hanekom[a)]
Bioengineering, Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Lynnwood Road, Pretoria 0002, South Africa
marianne.vanzyl@up.ac.za, johan.hanekom@up.ac.za

Abstract: The present study explored the acoustic characteristics of prosodic cues that indicate a speaker's reluctance when giving permission or agreement using a single word ("okay"). Eight speakers (four male, four female) produced the recorded materials that were subsequently validated through a listening experiment using 12 normal-hearing listeners. Acoustic analyses revealed that significantly longer word duration was the cue used most consistently across speakers to communicate reluctance. Voice quality, fundamental voice frequency, and intensity cues also differed significantly between the two prosodic conditions, but the manner in which these cues were applied varied greatly across speakers.

PACS numbers: 43.70.Fq, 43.71.Es, 43.72.Ar [AL]
Date Received: September 5, 2012 **Date Accepted:** November 15, 2012

1. Introduction

Knowing how speakers use prosodic cues to convey messages and the underlying acoustic cues will aid our understanding of an important communicative function and may be valuable in the assessment of amplification or spoken dialogue systems. Speech prosody is important in conveying both linguistic information such as stress or emphasis (Fry, 1955) and paralinguistic information such as emotions or attitudes (Murray and Arnott, 1993; Pell, 2007). Speakers often use prosody to convey communicative intentions, which are not always reflected by the content of their message (Sabbagh, 1999).

Prosody in single-word utterances can be used to fulfill linguistic functions such as marking syllable stress or differentiating questions from statements (Chatterjee and Peng, 2008; Fry, 1955) and paralinguistic functions such as communicating emotions (Hammerschmidt and Jürgens, 2007). Despite this, prosody of single-word utterances has received little attention in existing literature.

The present study focused on the use of single-word prosody to indicate a speaker's attitude, something that appears not to have been studied before. A particular realization of attitudinal prosody that can function on a single-word level was selected. Anecdotal evidence suggests that speakers may provide permission for or agreement with something although being reluctant. A reluctant attitude may be communicated through prosodic cues rather than the semantic content of the message, especially in a single-word response. Prosodic features such as a response delay and high boundary intonation are used to signal uncertainty in the context of answering a factual question (Brennan and Williams, 1995; Krahmer and Swerts, 2005). However, when answering a request, speakers may also use prosodic mechanisms to help communicate their attitude without necessarily revealing these attitudes in the content of the reply (Fujie *et al.*, 2006). The acoustic correlates of this phenomenon have not been reported in existing literature.

The present study examined prosodic cues to reluctance as expressed using the word "okay." This word was selected as a vehicle for this investigation as it is frequently

[a)]Author to whom correspondence should be addressed.

used for a wide variety of functions (Gaines, 2011) including agreement, and prosodic cues are important in differentiating its meaning in different contexts (Gravano *et al.*, 2012). In the present work, "okay" was used to communicate permission or agreement, and prosodic cues were used to communicate the attitude of the speaker (unreserved or reluctant). The objectives were (1) to confirm anecdotal evidence that normal-hearing listeners can detect a speaker's attitude from a single-word utterance such as "okay" and (2) to analyze the acoustic cues that differentiate these two attitudes.

2. Methods

2.1 Speech material development

Eight speakers (four male, MS1 to MS4; four female, FS1 to FS4) with normal hearing and speech, aged 21–30 yr, participated in the recording of speech material. Untrained speakers were used as the aim was to investigate the realization of the prosodic pattern in typical speakers. Fifty repetitions of the word "okay" were recorded from each speaker, 25 of which conveyed unreserved (baseline) permission and 25 conveying reluctant permission. To elicit these utterances, a scenario was sketched where someone would request to visit them on one of two different days. Speakers were informed that Friday would suit them in this scenario, whereas Monday would be inconvenient. Each elicited utterance was preceded by a question from the examiner (e.g. "Can I come on Monday?"), and speakers had to respond using only the word "okay," keeping in mind whether the requested time would be convenient or not. The same scenario was used to elicit all utterances, and this merits some explanation. Noting that "okay" performs a variety of communicative functions (Gaines, 2011), using the same scenario across elicitations ensured that the utterance was used to fulfill the same function in all instances. Also different scenarios could potentially induce a variety of emotions in the speakers, which had to be avoided. Baseline and reluctant elicitations were alternated to reduce task repetitiveness. Speakers were encouraged to produce each utterance as an authentic response to the examiner's question. Acoustic analyses of the recorded materials showed a high degree of variability within each speaker's collection of utterances, affirming that speakers were producing authentic responses rather than a rote repetition of the same utterance.

Speech material was recorded digitally in a double-walled sound booth with an M-Audio Fast Track Pro sound card (sampled at 44.1 kHz with 24-bit resolution) and a Sennheiser ME62 table-mounted microphone placed 20 cm from the speaker's mouth. Recorded speech materials were validated in 12 normal-hearing listeners (university students aged 19–29 yr) using a two-alternative forced-choice paradigm. No prior training was given to listeners, and no feedback was given during testing. This was to ensure that listeners would respond to the stimuli with everyday listening experiences as the only frame of reference. Different speakers' recordings were presented in counterbalanced order across listeners. Average scores for individual speakers across listeners varied between 72% and 95% (mean = 89%, standard deviation 7.38%). Utterances that were correctly classified by at least 10 of 12 listeners (i.e. significantly above chance, $p < 0.05$) were considered in the acoustic analyses.

2.2 Acoustic analyses

Acoustic characteristics were investigated using PRAAT (Boersma and Weenink, 2010) by examining aspects of fundamental voice frequency (F0), duration, intensity, and voice quality in each utterance. As a number of the distributions deviated significantly from a standard normal distribution, the non-parametric Mann–Whitney test was used to determine whether differences between the two conditions were significant ($p < 0.05$ or smaller). Effect size calculations were based on the Mann–Whitney test z-score and the total number of observations on which z is based (Field, 2009).

Average F0 and F0 range across the utterance were extracted, with assumed F0 ranges of 100–500 Hz for female speakers and 65–300 Hz for male speakers. An

average intonation contour of the final syllable was also determined for each speaker in both conditions. This required the elimination of duration differences between utterances without affecting their spectral characteristics, which was accomplished using phase vocoding methods (Ellis, 2002). The baseline and reluctant intonation curves for each speaker was then compared using Zhao's *z*-statistic for comparing trend curves (Zhao, 2011). Duration of the first syllable was measured from onset up to the end of the silence preceding the plosive noise of the /k/, and duration of the second syllable from the beginning of the release noise of /k/ to the end of phonation. Overall intensity (across the frequency spectrum) and voice quality of the voiced parts of the first and second syllables were determined separately. Voice quality was analyzed through extraction of the harmonics-to-noise ratio (HNR) with cross-correlation analysis in PRAAT. The HNR reflects the degree of periodicity in the utterance and consequently voice quality in areas where a valid F0 has been determined.

3. Results

Table 1 presents the values of each of the acoustic parameters in the validated recordings, and Table 2 shows effect sizes of the differences. Average F0 across the utterance was significantly higher in the baseline version for six of the eight speakers. Speaker FS4 produced a higher F0 average in the reluctant version, while speaker MS1 showed no significant difference between the F0 averages of baseline and reluctant versions. The F0 range (difference between maximum and minimum across the utterance) differed significantly between prosodic conditions for four speakers, three of which used a significantly greater range for reluctant prosody (FS1, FS2, MS2), whereas one speaker (FS3) produced a greater F0 range in the baseline condition. The intonation contours of the final syllable, as averaged over all the sampled utterances for each speaker, are shown in Figs. 1(a) (female speakers) and 1(b) (male speakers). Baseline and reluctant curves of each speaker were compared using a *z*-statistic (Zhao, 2011), with resulting p-values (Table 3) showing that five speakers produced intonation curves that differed significantly between the two versions. Table 3 also shows which half (first or last) of the utterances differed significantly.

All eight speakers used significantly longer total word and second syllable duration for reluctant utterances. The first syllable had a significantly greater duration in the reluctant versions of seven speakers. Except for MS3, the durational increase of the second syllable was greater than that of the first. MS3 used a longer first syllable, sometimes preceded by glottal fry or nasalization of the vowel, as a prominent cue of reluctant prosody. All the other speakers also produced audible aspiration noise at the end of most utterances, and this noise was significantly longer in reluctant versions of these speakers. Table 1 indicates the percentage of duration increase for each of the two syllables in the reluctant versions.

The intensity of the first syllable was significantly greater in the baseline version for five speakers, while the second syllable's intensity was significantly greater in the reluctant versions of six speakers. HNRs showed that voice quality differed significantly in one or both syllables for seven speakers with FS1 having a higher HNR in both syllables for baseline utterances, FS3 producing higher HNR in both syllables for reluctant utterances, and FS2, FS4, MS1, MS3, and MS4 produced higher HNRs for reluctant prosody on either the first or the second syllable.

Logistic regression analyses were carried out to explore the relative importance of the different cues in predicting to which category (baseline or reluctant) an utterance belonged. Different models were tested with predictors selected from the cues in Table 1. All validated utterances were included in the analyses. Utterances from male and female speakers were analyzed separately, as especially F0 parameters differed substantially between genders. Nagelkerke's R^2 was used as indicator for the variance accounted for. For both genders, all models that could account for more than 90% of the variance in the data set included duration as a predictor. Conversely, all models excluding duration as a predictor accounted for at most 66% of the variance. This confirms observations from Table 1 regarding duration being the most consistent cue. However, models that included

Table 1. Mean values of acoustic parameters for baseline (B) and reluctant (R) prosody. Female speakers are FS1–FS4; male speakers, MS1–MS4. For significant differences ($p < 0.05$ or smaller), the greater of the two values is indicated in boldface. S1 = 1st syllable; S2 = 2nd syllable.

Prosody	B	R	B	R	B	R	B	R
Speaker	FS1		FS2		FS3		FS4	
Number of utterances (*n*)	22	20	19	15	20	23	12	10
Average F0 (Hz)	**246.44**	214.23	**208.23**	193.61	**271.37**	225.10	260.78	**288.53**
F0 range (Hz)	127.31	**143.89**	94.20	**126.79**	**198.30**	97.85	194.22	209.44
Duration S1 (s)	0.19	**0.28**	0.10	**0.14**	0.15	**0.19**	0.16	0.17
Percentage increased duration	43.05		37.01		24.25		8.41	
Duration S2 (s)	0.27	**0.40**	0.23	**0.32**	0.27	**0.53**	0.27	**0.38**
Percentage increased duration	49.91		38.66		98.54		39.62	
Duration aspiration noise (s)	0.07	**0.20**	0.04	**0.09**	0.00	**0.03**	0.05	**0.11**
Total duration (s)	0.53	**0.88**	0.38	**0.54**	0.42	**0.75**	0.48	**0.66**
Intensity S1 (dB)	**74.30**	72.10	**71.74**	68.42	**72.76**	69.36	**74.36**	66.04
Intensity S2 (dB)	71.60	**72.57**	72.84	**73.49**	**72.72**	72.25	71.52	**72.98**
Intensity difference (S2-S1) (dB)	**−2.70**	0.47	1.10	**5.07**	−0.04	**2.89**	−2.84	**6.94**
Harmonics-to-noise ratio S1 (dB)	**15.81**	13.68	10.32	**13.85**	12.31	**13.99**	12.34	11.72
Harmonics-to-noise ratio S2 (dB)	**17.12**	13.84	14.69	14.98	15.70	**19.88**	12.72	**15.72**
Total significant differences	11/11		10/11		11/11		8/11	
Speaker	MS1		MS2		MS3		MS4	
Number of utterances	21	21	25	23	25	18	25	19
Average F0 (Hz)	118.81	109.34	**125.22**	99.57	**143.41**	133.55	**136.73**	117.77
F0 range (Hz)	77.37	80.82	55.41	**64.47**	61.57	74.85	61.18	60.13
Duration S1 (s)	0.19	**0.28**	0.16	**0.28**	0.19	**0.40**	0.16	**0.18**
Percentage increased duration	47.20		72.06		113.26		12.51	
Duration S2 (s)	0.27	**0.48**	0.14	**0.28**	0.21	**0.27**	0.21	**0.36**
Percentage increased duration	78.18		93.17		28.88		73.82	
Duration aspiration noise (s)	0.01	**0.05**	0.00	**0.06**	0.00	0.00	0.00	**0.03**
Total duration (s)	0.47	**0.82**	0.30	**0.62**	0.40	**0.67**	0.37	**0.56**
Intensity S1 (dB)	68.18	68.51	**72.29**	68.08	71.57	70.57	70.38	68.68
Intensity S2 (dB)	**73.31**	72.86	73.82	**74.73**	73.44	73.49	**73.29**	72.40
Intensity difference (S2-S1) (dB)	5.12	4.36	1.52	**6.65**	1.87	2.92	2.91	3.73
Harmonics-to-noise ratio S1 (dB)	7.76	6.74	6.87	6.44	8.76	**15.70**	7.05	7.13
Harmonics-to-noise ratio S2 (dB)	10.48	**12.57**	10.33	10.22	13.64	13.23	8.96	**11.39**
Total significant differences	6/11		9/11		5/11		7/11	

only duration as a predictor did not fully explain the data (male speakers, deviance = 47.9, degrees of freedom = 175, Nagelkerke's $R^2 = 0.895$; female speakers, deviance = 44.6, degrees of freedom = 139, Nagelkerke's $R^2 = 0.876$). Adding other predictors improved the models and evidence of cue trading relationships was observed. For example, for female speakers, models that included duration and either the intensity of both syllables (deviance = 24.4, degrees of freedom = 137, Nagelkerke's $R^2 = 0.937$) or average F0 and F0 range (deviance = 22.0, degrees of freedom = 136, Nagelkerke's $R^2 = 0.944$) did not differ significantly ($p = 0.124$).

4. Discussion

Perceptual validation confirmed that listeners were able to accurately discriminate between baseline and reluctant prosody in the recorded materials of all the speakers despite the inter-speaker differences in acoustic cues. The cue for reluctant prosody that was used with greatest consistency across speakers was an increase in duration.

Table 2. Effect sizes of differences between baseline and reluctant versions for each speaker (female speakers, FS1–FS4; male speakers, MS1–MS4). Effect sizes representing differences that were statistically significant ($p < 0.05$ or smaller) are depicted in boldface. S1 = 1st syllable; S2 = 2nd syllable.

Speaker:	FS1	FS2	FS3	FS4	MS1	MS2	MS3	MS4
Average F0 (Hz)	**−0.67**	**−0.60**	**−0.73**	**0.52**	−0.26	**−0.78**	**−0.45**	**−0.72**
F0 range (Hz)	**−0.31**	**−0.43**	**−0.60**	−0.15	−0.03	**−0.28**	−0.24	−0.01
Duration S1 (s)	**−0.84**	**−0.76**	**−0.43**	−0.30	**−0.83**	**−0.86**	**−0.84**	**−0.61**
Duration S2 (s)	**−0.86**	**−0.84**	**−0.86**	**−0.84**	**−0.86**	**−0.86**	**−0.78**	**−0.85**
Duration aspiration noise (s)	**−0.70**	**−0.64**	**−0.38**	**−0.65**	**−0.63**	**−0.68**	0.00	**−0.61**
Total duration (s)	**−0.84**	**−0.76**	**−0.86**	**−0.84**	**−0.86**	**−0.86**	**−0.85**	**−0.85**
Intensity S1 (dB)	**−0.56**	**−0.60**	**−0.65**	**−0.84**	−0.08	**−0.73**	−0.25	−0.26
Intensity S2 (dB)	**−0.35**	**−0.33**	**−0.42**	**−0.67**	**−0.50**	**−0.48**	−0.02	**−0.42**
Harmonics-to-noise ratio S1 (dB)	**−0.32**	**0.56**	**0.35**	−0.06	−0.17	−0.12	**0.70**	−0.04
Harmonics-to-noise ratio S2 (dB)	**−0.52**	−0.20	**−0.49**	**−0.67**	**−0.25**	−0.39	−0.14	**−0.57**
Average effect size	−0.60	−0.46	−0.51	−0.45	−0.45	−0.60	−0.29	−0.49

The importance of duration as a cue is confirmed by the effect sizes of the differences between prosodic versions and the amount of variance that this cue accounted for. Increased duration was also reported by Fujie *et al.* (2006) as an important cue of a negative response attitude, in addition to a smaller F0 range (which was not found to be a consistent cue in the present study), but the consistency of the cues across speakers was not reported.

Other cues were used less consistently, and the logistic regression analysis pointed to cue trading relationships. Observations regarding average F0 corroborate findings on other types of paralinguistic prosody such as sarcasm, where a reduction in F0 has been shown to be the most consistent prosodic cue (Cheang and Pell, 2008), and emotional prosody, where F0 changes constitute an essential acoustic cue (Williams and Stevens, 1972). Word final intonation has been reported to be important in the interpretation of the word "okay" in isolation (Gravano *et al.*, 2012). Some of the speakers in the present study used the intonation contour to differentiate baseline and reluctant attitudes, but different speakers applied intonation differently. Speakers FS1, MS2, and MS3 produced falling intonation contours in baseline utterances and rising

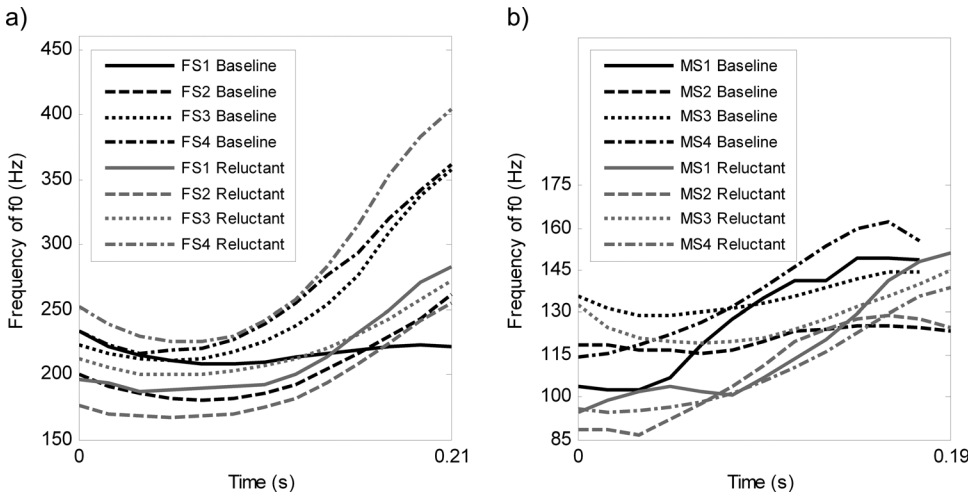

Fig. 1. Average intonation contours of final syllables as produced by female speakers numbered FS1 to FS4 (a), and male speakers numbered MS1 to MS4 (b), showing baseline and reluctant utterances separately.

Table 3. Results of the z statistic comparing the two utterance types (baseline and reluctant) of each speaker (female speakers, FS1–FS4; male speakers, MS1–MS4). Significant differences ($p < 0.05$) are depicted in boldface.

	Whole curve (p value)	1st half (p value)	2nd half (p value)
FS1	0.430	**0.010**	**0.010**
FS2	**0.005**	**<0.001**	0.213
FS3	**<0.001**	0.142	**<0.001**
FS4	**0.004**	**0.016**	**0.037**
MS1	0.052	0.156	0.134
MS2	**0.011**	**<0.001**	0.397
MS3	0.056	0.153	0.119
MS4	**<0.001**	**<0.001**	**<0.001**

contours in reluctant utterances, corresponding to findings regarding uncertainty in factual answers (Brennan and Williams, 1995), while the other speakers produced some form a rising pitch for both utterance types. Statistical comparison of the intonation curves showed that comparing the entire baseline curve with the entire reluctant curve may be useful in cases like speakers FS3 and MS4, where the two curves did not have any interaction, but may produce less informative results in cases like FS1, where the two curves clearly differed in shape (one rising and one falling or flat).

Speakers may use intensity as a cue to their attitude, but again the manner in which they apply this cue varies across speakers. Previous studies on cues for uncertainty in responses to factual questions did not report findings on intensity differences or values (Brennan and Williams, 1995; Krahmer and Swerts, 2005). Voice quality cues did not show consistent patterns across speakers, and effect sizes were small in comparison to most of the other investigated parameters. Higher HNRs observed in the reluctant versions of six of the speakers are in contrast to findings reported in a study on sarcasm, where a negative attitude corresponded to a lower HNR (Cheang and Pell, 2008).

In conclusion, (1) prosodic cues can differentiate unreserved and reluctant permission on the level of a single word, (2) the most consistent prosodic cue for distinguishing between reluctant and baseline single-word utterances was found to be duration, while cue trading between other cues were observed, and (3) the cues that communicate a baseline/reluctant attitude on a single word level are different than those that communicate emotion or that differentiate questions from statements. Previous work on acoustic characteristics of the word "okay" reported that word-final intonation, intensity, duration, mean F0, and voice quality all serve to differentiate different functions of the word (Gravano *et al.*, 2012), while the present work identified how these cues are applied to communicate the speaker's attitude and the consistency with which these cues are applied by different speakers. Future work could investigate whether these cues vary in different communication situations using the present data as a baseline.

Acknowledgments

The research was funded in part by the National Research Foundation of South Africa. The sponsor had no involvement in the design or implementation of the research or in the decision to submit the research report for publication.

References and links

Boersma, P., and Weenink, D. (**2010**). "PRAAT: Doing phonetics by computer (version 5.1.32) [computer program]," http://www.praat.org (Last viewed November 29, 2010).

Brennan, S. E., and Williams, M. (**1995**). "The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers," J. Mem. Lang. **34**, 383–398.

Chatterjee, M., and Peng, S. C. (**2008**). "Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition," Hear. Res. **235**, 143–156.
Cheang, H. S., and Pell, M. (**2008**). "The sound of sarcasm," Speech Comm. **50**, 366–381.
Ellis, D. P. W. (**2002**). "A phase vocoder in MATLAB," http://labrosa.ee.columbia.edu/matlab/pvoc (Last viewed November 26, 2010).
Field, A. (**2009**). *Discovering Statistics Using SPSS*, 3rd ed. (SAGE Publications Ltd., London).
Fry, D. B. (**1955**). "Duration and intensity as physical correlates of linguistic stress," J. Acoust. Soc. Am. **27**, 765–768.
Fujie, S., Ejiri, Y., Kikuchi, H., and Kobayashi, T. (**2006**). "Recognition of positive/negative attitude and its application to a spoken dialogue system," Syst. Comput. Jpn. **37**, 45–55.
Gaines, P. (**2011**). "The multifunctionality of discourse operator okay: Evidence from a police interview," J. Pragmat. **43**, 3291–3315.
Gravano, A., Hirschberg, J., and Benuš, S. (**2012**). "Affirmative cue words in task-oriented dialogue," Comput. Linguist. **38**, 1–39.
Hammerschmidt, K., and Jürgens, U. (**2007**). "Acoustical correlates of affective prosody," J. Voice **21**, 531–540.
Krahmer, E., and Swerts, M. (**2005**). "How children and adults produce and perceive uncertainty in audiovisual speech," Lang. Speech **48**, 29–53.
Murray, I. R., and Arnott, J. L. (**1993**). "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," J. Acoust. Soc. Am. **93**, 1097–1108.
Pell, M. D. (**2007**). "Reduced sensitivity to prosodic attitudes in adults with focal right hemisphere brain damage," Brain Lang. **101**, 64–79.
Sabbagh, M. A. (**1999**). "Communicative intentions and language: Evidence from right-hemisphere damage and autism," Brain Lang. **70**, 29–69.
Williams, C. E., and Stevens, K. N. (**1972**). "Emotions and speech: Some acoustical correlates," J. Acoust. Soc. Am. **52**, 1238–1250.
Zhao, Z. (**2011**). "Power of tests for comparing trend curves with application to national immunization survey (NIS)," Statist. Med. **30**, 531–540.