Identifying a creak probability threshold for an irregular pitch period detection algorithm

Olivia Murton,[1,a),b)] Stefanie Shattuck-Hufnagel,[2] Jeung-Yoon Choi,[2] and Daryush D. Mehta[1,b)]

[1]*Speech and Hearing Bioscience & Technology, Division of Medical Sciences, Harvard Medical School, Boston, Massachusetts 02115, USA*

[2]*Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

omurton@g.harvard.edu, sshuf@mit.edu, jyechoi@mit.edu, mehta.daryush@mgh.harvard.edu

Abstract: Irregular pitch periods (IPPs) are associated with grammatically, pragmatically, and clinically significant types of nonmodal phonation, but are challenging to identify. Automatic detection of IPPs is desirable because accurately hand-identifying IPPs is time-consuming and requires training. The authors evaluated an algorithm developed for creaky voice analysis to automatically identify IPPs in recordings of American English conversational speech. To determine a perceptually relevant threshold probability, frame-by-frame creak probabilities were compared to hand labels, yielding a threshold of approximately 0.02. These results indicate a generally good agreement between hand-labeled IPPs and automatic detection, calling for future work investigating effects of linguistic and prosodic context.

[BHS]
Date Received: April 8, 2019 **Date Accepted:** April 19, 2019

1. Introduction

In typical healthy speakers, non-modal phonation patterns occur in a wide variety of speech contexts, including at word boundaries, at phrase boundaries, and in certain prosodic contours. The type of non-modal phonation that is characterized by irregular pitch periods (IPPs) can also provide information about emotional content, dialect, speaker identity, and health status. IPPs exhibit varied acoustic realizations, including irregular spacing of pitch periods, single glottal pulses, and local decreases in *F*0 or amplitude.[1] Discussions of IPPs, creaky voice, and similar phonation types have often encountered terminological challenges. Researchers have used differing terms to describe these phenomena and have proposed a variety of theories about the mechanisms by which they occur and their functions in speech.[2,3] Here, we compare a strictly acoustic measure (the output of an algorithm) to labels created by human raters using a combination of visual and auditory-perceptual criteria.

In some languages, including American English, IPPs carry linguistic significance because of their association with lexical information, word or phrase boundaries, and prosodic prominences. Often corresponding to word boundaries, common locations for IPPs in American English include (1) the first few pitch periods of vowels or sonorant consonants at the beginning of an intonational phrase; (2) the last portion of an intonational phrase, particularly one with a low boundary tone; (3) the first few pitch periods at the onset of a high- or low-pitch accented syllable; and (4) the nucleus of a low-pitch-accented syllable where *F*0 is particularly low.[4] Additionally, in this dialect of English, IPPs can also convey lexical information by signaling a /t/ that occurs word-finally (e.g., *cat*) or between a stressed vowel and sonorant consonant (e.g., *butler*).[5]

IPPs are pragmatically significant when they are used to indicate speaker identity,[6] dialect,[7] and emotional state.[8] Speakers can vary widely from each other in their usage of IPPs, including both their baseline IPP usage across speech contexts and their grammatically-driven IPP usage in specific prosodic locations. Some speakers use IPPs more rarely than others, and speakers vary in their usage of IPPs to mark prosodic boundaries.[1] Although these differences present challenges for inter-speaker analyses of

[a)]Author to whom correspondence should be addressed.
[b)]Also at: Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, Boston, MA 02114, USA.

IPP usage, they also mean that IPPs may be useful for detecting unique aspects of specific speakers, like individual identity and dialect.

Finally, IPPs are clinically significant indicators of voice disorders[9] and systemic diseases including acute decompensated heart failure.[10] Holmberg *et al.*[9] used the perceptual term "scrape" to group together vocal fry, roughness, and irregular phonation. They found that patients with voice disorders presented with significantly less scrape after voice therapy compared to their pre-therapy baseline voices. Additionally, Murton *et al.*[10] used the detection algorithm discussed here to analyze the voices of patients undergoing treatment for acute heart failure. They found that these patients displayed a higher proportion of creaky voice after completing acute heart failure treatment, compared to their pre-treatment voices. IPPs are also relevant to automatic speech recognition and other systems that are aimed at extracting this information automatically.

In sum, episodes of IPP carry many different kinds of information, but despite their utility, identifying IPPs by hand is time-consuming and requires training. This difficulty makes reliable automatic detection of this phenomenon a compelling goal.[5] Previous work by Drugman *et al.* has used frame-based features, including short-term power, intra-frame periodicity, inter-pulse similarity, subharmonic energy, and glottal pulse peakiness, as inputs to an artificial neural network (ANN).[11–13] This ANN then assigns creak probabilities from 0 to 1 at regularly spaced time intervals across an acoustic recording. In contrast to the continuously varying creak probability, hand-labeling of IPPs or creaky voice is a binary decision—a frame is either in a creaky/IPP region or not. To be compared to hand labels, the algorithm's continuous output needs to be converted into a binary classification decision by identifying an appropriate creak probability threshold.

Drugman *et al.* performed this threshold identification process with a multilingual data set containing creaky voice hand labels. Their results indicated that creak probability thresholds that maximized the $F1$ score for each speaker were in the range of 0.3 for all speakers.[13] In this project, we perform a similar analysis with a data set that consists of conversations recorded by eight American English speakers and was hand-labeled for IPP regions by experienced labelers. Our goal is to determine whether the threshold identified in previous work is perceptually meaningful in our data set before we continue on to additional analyses based on IPP detection.

2. Methods

The American English Map Task corpus consists of acoustic recordings (16 kHz sampling rate) of 16 dyadic conversations by eight American English speakers who each wore a close-talk lapel microphone.[14] The speakers were all female, aged 18–22 years, and were familiar with each other. The recordings consisted of casual conversational speech during which one speaker (instruction giver) provided instruction to the other (instruction follower) to navigate through a map. Each speaker was an instruction giver in two conversations and instruction follower in two additional conversations. Only the data from instruction givers were included in the analysis. The data include four speakers recorded in two conversations, and four in a single conversation, for a total of 12 recordings.

For those 12 conversations, IPP regions were hand-labeled in Praat by trained raters following a standard labeling procedure similar to the one described in Dilley *et al.*[1] Raters identified speech regions with both (1) an auditory perception of non-modal voice quality associated with IPPs and (2) a visible irregularity in temporal spacing of periods in the acoustic waveform. Transcripts were also created by hand in Praat, indicating the time regions corresponding to each spoken word.

IPP regions were automatically identified using an algorithm developed for creaky voice analysis.[11–13] Window size settings were 25 ms for linear predictive coding analysis, 4 ms for short-term power, and 32 ms for intra-frame periodicity. Every 1 ms, the ANN generated a creak probability from 0 to 1, inclusive. Figure 1 illustrates the waveform, transcript, IPP hand labels, and creak probability contour for a short sample of the recorded speech. Only frames that occurred within a word were included; frames that occurred during silence, non-speech noise, or the instruction-follower's speech were discarded.

Because the outputs from the creaky voice algorithm varied continuously while the hand labels were binary, our goal was to choose a binary classification rule with the creak probability threshold that most closely aligned with the binary hand labels. Therefore, we used a variety of performance metrics to identify several different candidate thresholds and compared the classifier's performance at each threshold. We applied these performance metrics to a data set derived from concatenating all 12 conversations

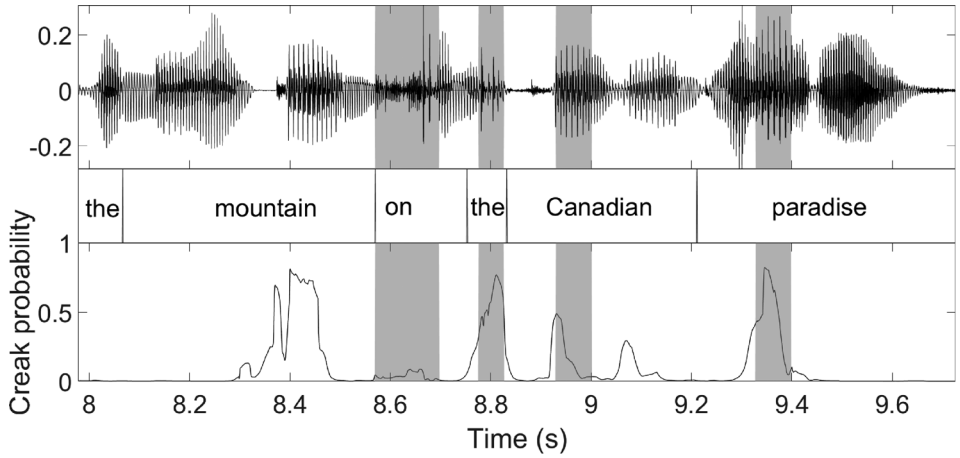

Fig. 1. Acoustic waveform (above) and automatically detected creak probability contour (below) from a section of Conversation 1 (Speaker 1). Shaded areas indicate hand-labeled IPP regions.

together. We swept thresholds from 0 to 1 and calculated the following performance metrics at each threshold: true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), positive predictive value (PPV), accuracy, and *F*1 score. The *F*1 score is based on the number of true positives, false negatives, and false positives. It is especially appropriate for evaluating performance when the positive class is comparatively rare in the data set, as in the case of laughter[15] or creaky voice/IPP usage.[13] We also plotted the FPR against TPR at each threshold to generate a receiver operating characteristic (ROC) curve, and calculated the area under the ROC curve (AUC) for the combination of all 12 conversations as well as each speaker's conversations individually. AUC is a measure of overall performance, where an AUC closer to 1 indicates better classification.

We compared creak probability thresholds across all speakers using four threshold criteria: (1) threshold at which TNR and TPR were equal (the equal error rate); (2) threshold yielding maximum *F*1 score; (3) threshold yielding maximum accuracy score; and (4) 0.3 creak probability, for comparison to results from Drugman *et al.*[13] Finally, to examine differences in algorithmic performance across speakers, we visualized the creak probability distributions of frames that were or were not hand-labeled as containing IPPs for each speaker.

3. Results

The instruction-givers' speech was recorded and labeled in 12 conversations, which totaled 98 min of recording. Approximately half of this recording time (48.5 min) consisted of the instruction-giver speaking. Approximately 11% (5.2 min) of this speaking time was hand-labeled as containing IPP segments, with a range of 6% to 18%. The variation in recording, speaking, and IPP region duration is reported in Table 1.

Figure 2 illustrates how the four creak probability threshold candidates were derived using the contours of the calculated performance metrics at each threshold. The intersection point of TNR and TPR gave the TPR = TNR threshold, and the location of the accuracy and *F*1 maxima gave the Max Accuracy and Max *F*1 thresholds. These contours were also used to obtain the values of multiple performance metrics at each candidate threshold. The threshold probabilities that this process identified, and

Table 1. Duration (in seconds) of total recording, speech segments (i.e., total time of all words as labeled in the Praat transcript), and IPP regions for each speaker who acted as an instruction giver. The percentage of speaking time that occurred in IPP regions is also reported.

Speaker	Recording time (s)	Speaking time (s)	IPP time (s)	IPP % of speaking time
1	726	301	32	11%
2	532	319	28	8.8%
3	524	296	26	8.8%
4	733	434	62	14%
5	953	496	29	5.9%
6	446	245	44	18%
7	1053	413	62	15%
8	894	405	30	7.5%

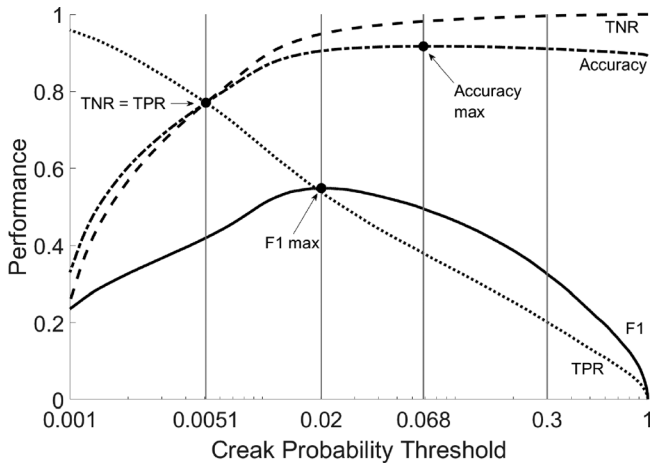

Fig. 2. Classifier performance in terms of various metrics (TNR, TPR, accuracy, and *F*1) for all conversations concatenated. The four creak probability threshold candidates (at TNR = TPR, max accuracy, max *F*1, and 0.3) are indicated with vertical lines. Note that the horizontal axis is logarithmic for visualization purposes.

the values of various performance metrics at each threshold, are presented in Table 2. The maximum *F*1 score of 0.549 was located at a threshold of 0.0202, which was used for additional analysis of individual speakers.

Using the TPR and FPR (given by 1 - TNR) in Fig. 2 yielded a ROC curve with an AUC of 0.853. The AUCs for equivalent ROC curves based on each speaker's conversations individually ranged from 0.797 to 0.926, with a mean of 0.864.

Overall, there were large differences in the distribution of creak probabilities for frames that were labeled as being in IPP regions compared to those that were not. However, the shapes of the probability distributions in IPP regions varied considerably by speaker, as shown in Fig. 3. Additionally, we calculated the *F*1 score for each speaker at the 0.0202 threshold, yielding a mean *F*1 of 0.539 (standard deviation of 0.084).

4. Discussion

Overall, we found that a lower threshold for creak probability than stated in the literature yielded satisfactory alignment with hand labels of IPP segments. While no single threshold maximized all performance metrics, our goal was to identify a threshold that best balanced the necessary trade-offs among different metrics. Because most speech frames were not creaky, it was important to choose a detection threshold carefully. Detection of rare events is more challenging than events that occur with more even probability, since the number of false positives can easily overwhelm the number of true positives if the threshold is set too low. However, setting the threshold too high risks missing many of the true positive tokens.

We chose the *F*1 score as the performance metric that best addressed this problem. The two higher thresholds (maximum accuracy and 0.3) under-identified creaky frames, leading to high PPV but low actual detection rates (TPR). In contrast, the lower threshold that equalized TPR and TNR had good detection rates, but the correspondingly high number of false positives made the PPV unacceptably low. A threshold located at the maximum *F*1 score gave high accuracy and identified over half of the hand-labeled creaky frames while still keeping the PPV above 0.5. Like Drugman *et al.*, we selected the maximum *F*1 score to find a threshold for detection of creaky frames in running speech. However, in our data set, the maximum *F*1 scores were typically achieved with a creak probability threshold of ~0.02, whereas Drugman

Table 2. Creak probability thresholds and performance metrics corresponding to four threshold criteria. Data are based on the group combining all speaker conversations.

Criterion	Creak threshold	*F*1	Accuracy	PPV	TPR	TNR
Max *F*1	0.0202	0.549	0.904	0.560	0.539	0.949
Max Accuracy	0.0684	0.496	0.917	0.713	0.340	0.981
TPR = TNR	0.0051	0.420	0.771	0.289	0.771	0.771
0.3	0.3	0.326	0.910	0.854	0.201	0.996

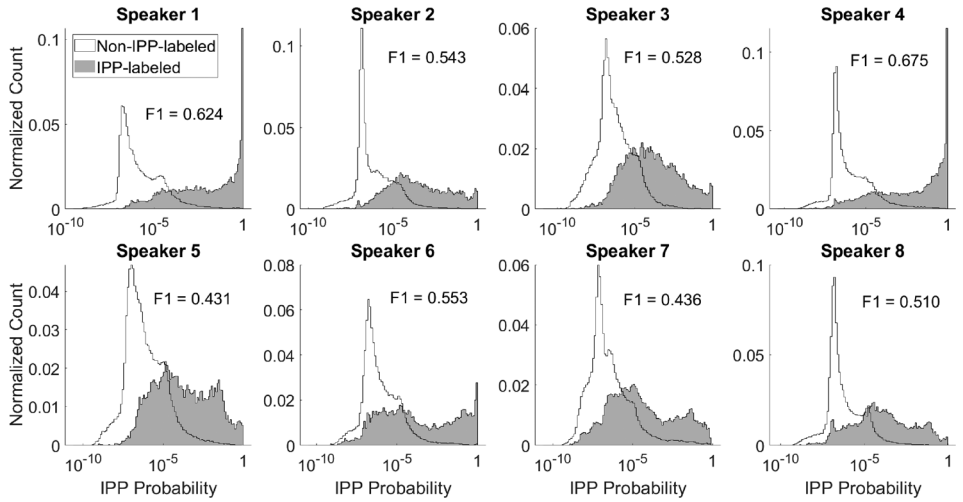

Fig. 3. Creak probability distributions of frames that were and were not hand-labeled as being in IPP regions for each speaker. Total bin counts are normalized to sum to 1 within each IPP label (IPP or non-IPP) for each speaker. Text indicates each speaker's *F*1 score at the 0.0202 threshold, which gave the maximum *F*1 score for all conversations combined.

et al. found their *F*1 maxima using a creak probability threshold of ~0.3.[13] Further, the detection output is intended to indicate the probability of a given frame lying in a creaky/IPP region. Therefore, a naively chosen threshold probability would be 0.5, i.e., a frame would be considered to fall in an IPP region if the probability assigned by the detector were above chance. Here, instead, we find that frames assigned a probability of only 0.02 aligned well with hand-labeled IPP regions; in other words, that a frame with just a 3% "probability" of falling in an IPP region can actually be expected to do so. A threshold of 0.02 is so low that it is effectively meaningless as a probability of creak/IPP. This unexpected finding suggests that the algorithm's output may be reflecting some underlying phenomenon that is different from creaky or IPP phonation, and calls for further investigation.

Although the detection algorithm performed well at the lower threshold, the large difference between our results and those from Drugman *et al.* suggests that our input data set are different from theirs in some important way. That difference is likely related to our hand-labeling techniques and/or our speaker populations. The hand-labeling process of identifying creaky regions both auditorily and visually is similar to the process described by Drugman *et al.*, but many factors could have caused differences in output despite that broad similarity. Our speakers were all American English speakers, while Drugman *et al.* analyzed data sets of speakers speaking Swedish, Finnish, and Japanese in addition to American English. The acoustic realizations of IPPs may differ across languages and cultures, causing the algorithm to interact differently with our speakers than with speakers of other languages.

For example, 11% of speech in our data set was hand-labeled as being in an IPP region, compared to approximately 6% of speech in the Drugman corpora. Speakers of American English in the Drugman corpus tended to have higher proportions of creaky/IPP-labeled speech (7.7%) compared to non-English speakers (5.7%), which may explain why our proportion of IPP-labeled speech was higher. Like previous studies on different corpora,[1] our own results (Fig. 3) indicate that even speakers within a demographic group (in this case, young female American English speakers) show different distributions of detection probability in speech frames that are labeled in IPP regions. Our results indicate that group-specific patterns of IPP usage and acoustic production may affect the performance of this automatic detection algorithm and should be considered when applying it to novel speaker populations.

The results presented here raise a number of questions for future investigation. First, these results are based on frame-by-frame comparisons of hand and automatic labels. However, IPP regions typically span many frames, so this frame-by-frame comparison may be unnecessarily strict. For example, if the hand labelers and algorithm identified 1-s IPP regions with 80% overlap, it might be sufficient to treat the algorithm as having detected that IPP region, without penalizing it for the false positive and false negative frames on the region's boundaries. This approach would require determining

a minimum overlap amount necessary for "detection" of an IPP region, as well as an appropriate probability threshold for the classifier. Taking this more generous event-based approach to IPP detection might give a more accurate picture of the agreement between human and automatic labels.

Second, as discussed previously, IPPs can have varied acoustic realizations across speakers and speech contexts. Drugman *et al.* identified three acoustic patterns of creaky voice production and reported different detection rates for each pattern.[13] Examining our data in light of these or other patterns might help explain our classifier's performance in different speech contexts, across speakers, and in comparison to other experimental results. For example, our data set is also labeled for prosodic structure with Tones and Break Indices labels that capture phrasing and prominence.[16] Using automatic detection results to distinguish different acoustic realizations of IPPs could expand on previous work regarding the interactions between IPP realization and prosodic location[1–3] and help to better understand the algorithm's performance in a variety of prosodic contexts. Additionally, because we classified frames with an output probability above 0.02 as IPP frames, our data set contains IPP frames with a very wide range of assigned probabilities. It is possible that different IPP probability ranges correspond to distinct acoustic realizations of IPP, or have other meaningful differences between them.

Finally, IPP-like voice characteristics are known to interact with voice disorders[9] and systemic disease.[10] Applying this algorithm to clinical populations could provide more information about how these disorders affect IPP production.

5. Conclusion

IPPs in speech are informative, but challenging to detect because they require time and training to label by hand. We investigated appropriate detection thresholds for an algorithm that uses multiple acoustic features input to an ANN to automatically yield creak probabilities that were hypothesized to be related to IPP behavior. This work has implications for use in future work investigating different productions and realizations of IPPs, relating IPP usage to specific prosodic contexts, and understanding the relationships between IPP usage and human health states.

Acknowledgments

This project is supported by grants from the NIH National Institute on Deafness and Other Communication Disorders (Grant Nos. T32 DC000038, P50 DC015446, R21 DC015877), the NSF Division of Behavioral and Cognitive Sciences (Grant No. 1651190), and the Center for Assessment Technology and Continuous Health (CATCH) at Massachusetts General Hospital. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

References and links

[1]L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of word-initial vowels as a function of prosodic structure," J. Phonetics **24**(4), 423–444 (1996).

[2]B. R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," J. Phonetics **29**(4), 365–381 (2001).

[3]P. A. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *Proceedings of the International Congress of Phonetic Sciences*.

[4]J. Pierrehumbert and D. Talkin, "Lenition of /h/ and glottal stop," in *Papers in Laboratory Phonology II*, edited by G. Docherty and D. Ladd (Cambridge University Press, Cambridge, UK, 1992), pp. 90–116.

[5]L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," J. Phonetics **29**(4), 407–429 (2001).

[6]T. Bőhm and S. Shattuck-Hufnagel, "Do listeners store in memory a speaker's habitual utterance-final phonation type?," Phonetica **66**(3), 150–168 (2009).

[7]C. G. Henton, "Creak as a sociophonetic marker," J. Acoust. Soc. Am. **80**(S1), S50 (1986).

[8]C. Gobl, "The role of voice quality in communicating emotion, mood and attitude," Speech Commun. **40**(1–2), 189–212 (2003).

[9]E. B. Holmberg, R. E. Hillman, B. Hammarberg, M. Södersten, and P. Doyle, "Efficacy of a behaviorally based voice therapy protocol for vocal nodules," J. Voice **15**(3), 395–412 (2001).

[10]O. Murton, R. Hillman, D. Mehta, M. Daher, T. Cunningham, K. Verkouw, S. Tabtabai, J. Steiner, G. W. Dec, D. Ausiello, and M. Semigran, "Acoustic speech analysis of patients with decompensated heart failure: A pilot study," J. Acoust. Soc. Am. **142**(4), EL401–EL407 (2017).

[11]J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," Comp. Speech Lang. **27**(4), 1028–1047 (2013).

[12]C. T. Ishi, K.-I. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," IEEE Trans. Audio, Speech, Lang. Process. **16**(1), 47–56 (2008).

[13]T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," Comput. Speech Lang. **28**(5), 1233–1253 (2014).
[14]README_AmEngMapTask. Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/32533/README_AmEngMapTask.pdf (Last viewed 3/12/2019).
[15]S. Scherer, F., Schwenker, N., Campbell, and G. Palm, "Multimodal laughter detection in natural discourses," in *Human Centered Robot Systems* (Springer, Berlin, Heidelberg, 2009), pp. 111–120.
[16]M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, *Prosodic Typology*, edited by S.-A. Jun (Oxford University Press, Oxford, UK, 2005).