DETECTION OF QUESTIONS IN CHINESE CONVERSATIONAL SPEECH

Jiahong Yuan & Dan Jurafsky

Stanford University
{jy55, jurafsky}@stanford.edu

ABSTRACT

What features are helpful for Chinese question detection? Which of them are more important? What are the differences between Chinese and English regarding feature importance? We study these questions by building question detectors for Chinese and English conversational speech, and performing analytic studies and feature selection experiments. As in English, we find that both textual and prosodic features are helpful for Chinese question detection. Among textual features, word identities, especially the utterance-final word, are more useful than the global (N-gram) sentence likelihood. Unlike in English, where final pitch rise is a good cue for questions, we find in Chinese that utterance final pitch behavior is not a good feature. Instead, the most useful prosodic feature is the spectral balance, i.e., the distribution of energy over the frequency spectrum, of the final syllable. We also find effects of tone, e.g., that treating interjection words as having a special tone is helpful. Our final classifier achieves an error rate of 14.9% with respect to a 50% chance-level rate.

1. INTRODUCTION

Identification of dialogue acts (DAs), such as statements, questions, backchannels, and so on, is of fundamental importance to automatic understanding of natural speech. In this study, we investigate automatic question detection, a particular task of DA identification, in conversational Mandarin Chinese speech.

Both word-based and acoustic-prosodic information have been used for English DA identification. There have been two approaches to using word-based information: N-gram language models and Word Identity. In the N-gram approach [1,2,3], the transcripts of the training corpus are grouped by DA type, and an N-gram LM is trained for each DA type. The LMs are used to compute a likelihood of the entire word sequence for each DA type, and the DA whose LM assigns the maximum likelihood is chosen. In the Word Identity approach [3,4,5], particular words are extracted and used as features in classification, for example, the first and last word of an utterance, cue words and phrases [6]. A performance comparison of the two approaches has not been reported in the literature.

Although words are the primary cue for DA identification, acoustic-prosodic information is also helpful [1,7]. Furthermore, there are motivations for investigating DA identification from prosodic cues alone [1,8], such as the high error rate of conversational recognizers and the fact that e.g., some questions have the same word order as statements (e.g., "tomorrow?", "He is a student?"), and hence can only be identified via prosody. Prosodic features based on pitch, energy, and duration have been used by a number of labs for DA identification, and the relative importance of the features discussed, for example in [1].

Chinese is very different from English and other Indo-European languages, which most DA identification studies focused on. First, while English questions often have 'inverted' word order from statements, Chinese questions have the same word order as statements. The particle '吗 (ma)' can be added to any statement and convert it into a yes-no question. There is also a special type of question called the A-not-A question [9], which is formed with the main verb followed by negation '不(bu)' or '没(mei)' and the reduplicated verb. Listed below are examples of common types of questions in Mandarin Chinese:

Statement:	他 想 去.	
	he want go	
	"He wants to go."	
Yes-no question:	他 想 去 吗?	
	he want go ma	
	"Does he want to go?"	
Echo question:	他 想 去?	
	he want go	
	"He wants to go?"	
A-not-A question:	他 想 不 想 去?	
	He want not want go	
	"Does he want to go or not?"	
Wh-question:	他 为什么 想 去?	
	he why want go	
	"Why does he want to go?"	

We hypothesize that cues to Chinese questions may be extremely local, and hence that specific word identities may function better than the total sentence likelihood assigned by the N-gram model. Furthermore, since question markers are at the end of sentences, sentence-final words should be more helpful than sentence-initial words in Chinese question detection. We'll test these hypotheses and compare the results with English.

Second, Chinese is a tonal language. The difference between question and statement intonation in Chinese is complicated because of the interaction of tone and intonation. For example, interrogative intonation on a final word with a rising tone has a rising end, resembling the final rise in English, whereas question intonation on a final word with a falling tone often has a falling end (as shown in the last 100 ms of Figure 1).

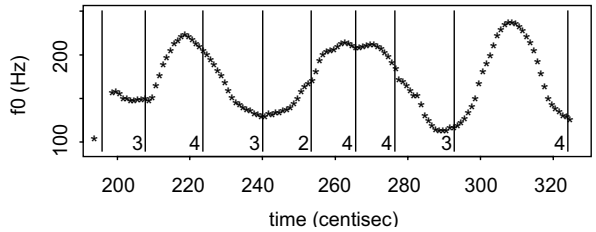

Figure 1: Interrogative intonation in Chinese can have a falling tail. Vertical lines mark syllable boundaries. Numbers indicate tones.

Perceptual studies have shown that the final tones affect intonation perception. Question intonation, for example, is easier to recognize by a human if the sentence-final tone is falling whereas it is harder to recognize if the sentence-final tone is rising [10]. How, then, do tones affect automatic question detection? What prosodic features are more useful? What is the difference between Chinese and English on the importance of particular prosodic features? We did experiments to answer these questions.

The paper is organized as follows: in the next section we present the data used for our experiments. In section 3, we discuss the features useful for question detection in Chinese and their relative importance. In section 4, we report the results of our evaluation. In section 5, we compare the difference between Chinese and English on feature importance. Finally, we present our conclusions.

2. DATA

Our data were taken from the CALLHOME Mandarin Chinese corpus of telephone speech (LDC96S24) and its transcripts (LDC96T16). The transcripts are timestamped by speaker turn. We selected all the speaker turns that conclude with either a period or a question mark in the transcripts. These turns were then word aligned using the SONIC speech recognition system [11], with an acoustic model trained on the CALLHOME and CALLFRIEND (LDC96S55) Mandarin Chinese corpora. Since a turn may contain more than one dialogue act, we extracted and used only the last part of each selected turn, that is, from where the second to the end punctuation is, or from the beginning if there is only an ending punctuation, to the end of the turn. The data sample was tagged as a statement if ending with a period and a question if ending with a question mark.

There were more statements (75.4%) than questions (24.6%) in the data. To avoid the imbalanced data problem, we downsampled the statements to obtain an equal number of statements and questions, following [1]. Our training set contained 3085 statements and 3085 questions, and dev set contains 642 statements and 642 questions. The training and dev sets were partitioned in the CALLHOME corpus. They shared no speakers. Below are some examples taken from the training set:

Statements	Questions
我家里电话变了.	那你现在都还好吗?
随便注册一个什么公司.	那盆小的?
对.	你到底上不上?
哦.	哦?

3. FEATURES

In this section, we study the roles of textual features, prosodic features, and tones in question detection. All experiments were run using the decision tree classifier C4.5 [12] and doing 5-fold cross validation on the training set. The add-one-in technique was used for doing feature selection.

3.1. Textual features

To compare the performance of the N-gram language model approach and the Word Identity approach, and to find if the classifier can be improved by combining the two approaches, we calculated the difference between the log-likelihoods of a sentence (a word sequence) being a question and being a statement, and used it as a textual feature of the sentence. The other textual features included word identities, including the first and last word of the sentence, the number of words in the sentence, whether there is a Wh-word, A-not-A construction, negation, and the word 'you' or 'your' in the sentence.

Table 1. Textual features selected by add-one-in.

Feature added	Error rate
Last word	18.4%
+ Wh-word	17.6%
+ A-not-A	17.0%
+ The number of words	17.0%
+ 'you' or 'your'	16.9%
Using N-gram sentence probability only	21.7%

We did feature selection on these features. The results are listed in Table 1. We can see that the N-gram sentence probability feature was not selected; and the error rate of using the N-gram sentence probability feature (21.7%) was about 5% (absolute) higher than that of using the selected feature vector (16.9%). From these results we conclude that word identities are more useful than the global (N-gram) sentence likelihood for Chinese question detection. The last word was the first selected feature whereas the first word was excluded, showing that as a textual feature the sentence-final word is more useful than the sentence-initial word.

3.2. Prosodic features

3.2.1. Normalization

Speech prosody manifests itself in pitch, energy, and duration. There is significant variation among speakers in these acoustic measures. To normalize prosodic features by speakers, we first located the maximum and minimum pitch, energy, and syllable duration of each speaker, after exclusion of the speaker's highest and lowest 2% data to eliminate measurement and alignment errors; then we normalized the raw data by setting the maximum as 10 and the minimum as 0, as shown in the equation:

$$X_{Norm} = (X_{Raw} - Min) \times \frac{10}{(Max - Min)}$$

The raw F_0 data (in Hz) were logarithmized before normalization. The raw energy (in dB) and duration (in second) data were directly used for normalization.

3.2.2. Pitch features

Many previous studies point to the importance of the utterance-final region to the realization of question intonation. The utterance final behavior of question intonation has been interpreted differently in Chinese intonation models, including via boundary tone [13], prosodic strength [14], baseline and topline [15]. Guided by these models, we extracted four pitch features from the utterance final syllable: the highest pitch over the final syllable, the lowest pitch over the final syllable, the pitch range of the final syllable, and the pitch at the end of the final syllable.

Besides utterance final behavior, previous studies also found that question intonation has an effect on the pitch of the whole utterance [14,16,17]. To capture this effect, we fitted a linear regression line to the pitch curve of each utterance, and used the slope and the intercept of the fitted line as two more pitch features, encoding the direction and the height of the pitch curve.

3.2.3. Energy features

Compared to pitch, energy has drawn much less attention in the literature of dialogue act realization. Tsao claimed that question intonation in Chinese is 'a matter of stress' [18]. There are also quantitative studies showing that in laboratory speech the overall intensity of sentence final syllables is greater in questions than in statements [19].

On the other hand, the role of energy in accent and stress realization has been extensively studied. Many studies show that spectral balance (or spectral emphasis, spectral tilt), measuring the distribution of energy over the frequency spectrum, is a more reliable acoustic cue of accent and stress than the overall intensity [20]. Spectral balance has also been used as a feature in pitch accent detection [21,22] and disfluency identification [23], and proved more useful than the overall intensity in pitch accent detection [21].

To study the role of spectral balance in question detection and to compare it with overall intensity, we did feature selection on the following energy features, which were extracted over the last syllable:

Int_overall: the overall intensity;
Int_0_05: the intensity in the frequency band of 0-0.5 kHz;
Int_05_1: the intensity in the frequency band of 0.5-1 kHz;
Int_1_2: the intensity in the frequency band of 1-2 kHz;
Int_2_4: the intensity in the frequency band of 2-4 kHz;
Int_balance: the difference between *Int_1_2* and *Int_0_05*.

The results are listed in Table 2.

Table 2. Energy features selected by add-one-in.

Feature added	Error rate
Int_balance	36.0%
+ *Int_05_1*	35.9%
+ *Int_2_4*	35.7%

Spectral balance was first selected whereas the overall intensity was excluded from the final feature set, showing that, as in English accent detection, spectral balance is more useful than the overall intensity in Chinese question detection. The intensity over 0.5 to 1 kHz and over 2 to 4 kHz were also selected, although they only contributed to a small error rate reduction (0.3%).

3.2.4. Duration features

The duration difference between statement and question intonation in Chinese has been studied in [19]. In general, the utterance final syllable in question intonation tends to be longer than in statement intonation; whereas the other syllables in question intonation tend to be shorter. Therefore, we extracted three duration features: the duration of the final syllable, the average duration of the other syllables, and the length of the whole utterance (the last feature was not normalized).

3.2.5. Feature selection

We've introduced three groups of prosodic features. They are listed in Table 3.

Table 3. Prosodic features.

Pitch features	*Slope*: the slope of the fitted line;
	Intercept: the intercept of the fitted line;
	Max: the highest pitch over the last syllable;
	Min: the lowest pitch over the last syllable;
	Range: the pitch range of the last syllable;
	End: the pitch at the end of the final syllable;
Energy features	*Int_balance*: *Int_1_2* minus *Int_0_05*;
	Int_05_1: the intensity over 0.5-1 kHz;
	Int_2_4: the intensity over 2-4 kHz;
Duration features	*Last*: the duration of the final syllable;
	Pre: the average duration of the previous syllables (other than the last syllable);
	Length: the length of the whole utterance;

To explore their relative importance, we did feature selection on the prosodic features. The results are listed in Table 4.

Table 4. Prosodic features selected by add-one-in.

Feature added	Error rate
Int_balance (energy)	36.0%
+*Length* (duration)	34.0%
+*Slope* (pitch)	31.9%
+*Last* (duration)	30.6%
+*Intercept* (pitch)	30.5%

Used in isolation, the spectral balance of the final syllable is the most useful prosodic feature. None of the utterance-final pitch features, on the other hand, were selected by add-one-in. This result suggests that at the utterance-final position energy is more reliable than pitch as a prosodic cue of question intonation in Chinese. Whether boundary tone or prosodic strength better explains question intonation in Chinese is an open problem. Our result favors prosodic strength over boundary tone.

The two global pitch features, representing the direction and the height of the overall pitch curve, and two duration features, the length of the utterance and the duration of the last syllable, were also selected, which is consistent with phonetic studies of Chinese intonation.

3.3. Tones

There are four citation tones and a neutral tone in Mandarin Chinese, referred to as Tone1, Tone2, Tone3, Tone4, and Tone0. We found that the classification error rate was up about 4%, from 30.5% to 34.3% when excluding the final tone from the prosodic feature vector. This result shows that tonal information is helpful in Chinese question detection.

There are a great number of utterances in our data, both statements and questions, containing only an interjection word. In Chinese dictionaries, the interjection words usually have more than one tone. For example, in the *Modern Chinese Dictionary*, '哦 (oh)' is listed with two tones, tone2 (a rising tone), indicating 'not totally believe something' and tone4 (a falling tone), indicating 'got it'. Obviously, the intonation forms on the interjection word were interpreted as its tones. Figure 2 draws the (normalized) pitch curves of the interjection words appeared in a one-word statement or question (randomly selected 300 statements and questions were drawn).

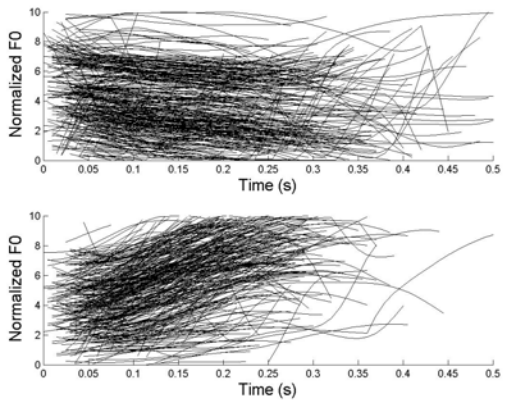

Figure 2: The (normalized) pitch curves of interjections in statements (shown on the top) and questions (shown on the bottom).

The interjections have a falling pitch curve in statements and a rising pitch curve in questions, suggesting that they are toneless. We may either treat the interjections as having a neutral tone, or as having a different tone, which we could call Tone5. From Table 5, we can see that treating the interjections as a special tone slightly improve the performance of the prosodic features.

Table 5. Error rates of using tones in different ways.

Tonal categories	Error rate
Tonal information excluded	34.3%
Interjections as neutral tone (5 tones)	30.5%
Interjections as a special tone (6 tones)	29.7%

4. EVALUATION

We discussed above the textual and prosodic features useful for Chinese question detection, and their relative importance. In this section, we evaluate these features on the withheld dev set. The results are listed in Table 6.

Table 6. Error rates of using different feature vectors, trained on the training set and tested on the dev set.

Features		Error rate
Textual features	N-gram sentence probability	19.8%
	Word identities	16.1%
Prosodic features	Tonal information excluded	34.4%
	Interjections as a neutral tone	29.0%
	Interjections as a special tone	27.0%
Textual and prosodic features		14.9%

The results shown in Table 6 are consistent with what we have found in the previous section. First, word identities are more useful than the N-gram sentence probability (16.1 % vs. 19.8 %; $p < .01$ by McNemar's test); secondly, tonal information helps on the performance of the prosodic features (29.0% vs. 34.4%; $p < .0001$); thirdly, treating the interjections as a special tone is beneficial (27.0% vs. 29.0%; $p < .0001$).

We also trained a classifier using both the textual and the prosodic features (feature selection was done on the training set). The error rate of the classifier was 14.9%, a 1.2% reduction from using the textual features only. The reduction is statistically significant ($p < .05$).

5. COMPARISON WITH ENGLISH

Our study on Chinese question detection made two interesting findings: first, the word identity features are more useful than the N-gram sentence probability, and the utterance final word is the most useful textual feature; second, the spectral balance of the final syllable is the most useful prosodic feature, while utterance final pitch features, don't help. Previous studies on English question detection have not directly compared the performance of using word identities and using N-gram sentence probability; and spectral balance has not been used as a feature for English question detection. In this section we report our experiments on English and make a comparison between Chinese and English.

5.1. Data

The English data were taken from the CALLHOME English corpus of telephone speech (LDC97S42) and its transcripts (LDC97T14), using the same method as obtaining the Chinese data. A considerable portion of the English CALLHOME transcripts doesn't contain punctuations, and hence are not usable for this study. We combined the training set and dev set of the English data. The combined data set has 3172 datapoints, including 1586 statements and 1586 questions. We did 5 fold cross validation on the combined data set in the following experiments.

5.2. Textual features

We did feature selection on the same textual features as used for Chinese, including both N-gram probability and word identities. The results are listed in Table 7.

Table 7. Textual features selected by add-one-in (English)

Feature added	Error rate
First word	26.2%
+ 'you' or 'your'	23.8%
+ Wh-word	23.3%
+ The number of words	22.5%
Using N-gram sentence probability only	29.0%

Compared to Chinese, the error rates are considerably higher. This is probably because we have a smaller data set for English.

The word identity features are more helpful than the N-gram sentence probability, just as we found for Chinese. The most important textual feature, however, is the first word instead of the last word. This is an expected result because in English the yes-no questions start with auxiliary verbs, whereas in Chinese the yes-no questions end with a question marker.

5.3. Prosodic features

Table 8 lists the results of the prosodic feature selection, starting with the features in Table 3 (the features were extracted over words for English), plus the overall intensity of the last word.

Table 8. Prosodic features selected by add-one-in (English)

Feature added	Error rate
End (Pitch)	36.5%
+*Length* (Duration)	35.8%
+*Last* (Duration)	35.0%
+*Int_balance* (Energy)	34.3%
+*Min* (Pitch)	34.2%

In English, quite differently than in Chinese, the pitch at the end of the final word was first selected. This result is consistent with the ToBI model of English intonation [24], in which question intonation has a high boundary tone. The difference between English and Chinese suggests, on the other hand, that the boundary tone model is not suitable for Chinese question intonation.

Both the overall intensity and the spectral balance were included for feature selection, but only the spectral balance was selected. This confirms that spectral balance is a more reliable energy feature for question detection.

It is the spectral balance but not the utterance final pitch features that were selected in both Chinese and English. This result suggests that the language universal effect of

intonation is probably more related to energy than to pitch. Further research is needed in this direction.

6. CONCLUSIONS

We investigated the features useful for Chinese question detection and their relative importance. Our classifier achieves an error rate of 14.9% with respect to a 50% chance-level rate. We also compared the differences between Chinese and English regarding feature importance in question detection. We made the following conclusions:

Specific features related to word identity are more useful than using a full-sentence N-gram probability for both Chinese and English question detection. The utterance-final word is the most useful textual feature for Chinese whereas the utterance-initial word is most useful for English.

For question detection spectral balance is a more reliable energy feature than the overall intensity. The spectral balance of the final syllable is the most useful prosodic feature for Chinese; the pitch at the end of the utterance, however, is the most useful prosodic feature for English. This result suggests that boundary tone is a suitable intonation model for English but not for Chinese.

Tonal information helps Chinese question detection. It is also helpful to categorize the interjection words as a special tone.

7. ACKNOWLEDGEMENT

We thank Prof. Jiong Shen for suggesting the spectral balance feature, and thank Prof. Bryan Pellom for helping on using SONIC. This research was partly funded by the Edinburgh-Stanford LINK and by the ONR.

8. REFERENCES

[1] Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., Van Ess-Dykema, C., **"**Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" *Language and Speech*, 41(3-4), 439-487, 1998.

[2] Stolcke, A., Ries, k., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Meteer, M., and Van Ess-Dykema, C., "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," *Computational Linguistics* 26(3), 339-371, 2000.

[3] Hillard, D., Ostendorf, M., and Shriberg, E., "Detection Of Agreement vs. Disagreement In Meetings: Training With Unlabeled Data," *Proceedings of HLT-NAACL*, Edmonton, Canada, 2003.

[4] Ang, J., Liu, Y., and Shriberg, E., "Automatic Dialog Act Segmentation and Classification in Multiparty Meetings," *Proceedings of ICASSP*, Philadelphia, 2005.

[5] Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E., "Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies," *proceedings of ACL-04*, Barcelona, Spain, 2004.

[6] Hirschberg, J., and Litman, D., "Empirical Studies on the Disambiguation of Cue Phrases," *Computational Linguistics*, 19(3), 501-530, 1993.

[7] Mast, M., Kompe, R., Harbeck, S., Kieling, A., Niemann, H., Noth, E., Schukat-Talamazzini, E.G., and Warnke, V., "Dialog Act Classification with the Help of Prosody," *Proceedings of ICSLP 1996*, Philadelphia, 1996.

[8] Fernandez, R., and Picard, R.W., "Dialog Act Classification from Prosodic Features Using Support Vector Machines," *Speech Prosody 2002*, Aix-en-Provence, France, 2002.

[9] Li., C., and Thompson S., *Mandarin Chinese: A Functional Reference Grammar*, Univ. of California Press, Berkeley, 1981.

[10]Yuan, J., and Shih, C., "Confusability of Chinese Intonation," *Speech Prosody 2004*, 131-134, Nara, Japan, 2004.

[11]Pellom, B., "SONIC: The University of Colorado Continuous Speech Recognizer," University of Colorado, tech report #TR-CSLR-2001-01, Boulder, Colorado, 2001.

[12]Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, 1993.

[13]Peng, S., Chan, M., Tseng, C., Huang, T., Lee, O., and Beckman, M.E., "Towards a Pan-Mandarin system for prosodic transcription," In: Sun-Ah Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*, pp. 230-270, Oxford University Press, Oxford, U.K., 2005.

[14]Yuan, J., Shih, C., Kochanski, G.P., "Comparison of Declarative and Interrogative Intonation in Chinese," *Speech Prosody 2002*, 711-714, Aix-en-Provence, France, 2002.

[15]Shen., J., "Hanyu yudiao gouzao he yudiao leixing [Intonation structure and intonation types of Chinese]," *Fangyan*, 3, 221-228, 1994.

[16]Shen, X., *The Prosody of Mandarin Chinese,* University of California Press, Berkeley, 1989.

[17]Garding, E., "Speech Act and Tonal Pattern in Standard Chinese: Constancy and Variation," *Phonetica*, 44, 13-29, 1987.

[18]Tsao, W., "Question in Chinese," *Journal of Chinese Language Teachers' Association*, 2, 15-26, 1967.

[19]Yuan, J., *Intonation in Mandarin Chinese: Acoustics, Perception, and Computational Modeling*, Ph.D. Dissertation, Cornell University, Ithaca, 2004.

[20]Sluijter, A., and Van Heuven, V., "Spectral balance as an acoustic correlate of linguistic stress," *JASA*, 100(4), 2471-2485, 1996.

[21]Heldner, M., "Spectral Emphasis as an Additional Source of Information in Accent Detection," *Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, 57-60, Red Bank, NJ. 2001.

[22]Zhang, T., Hasegawa-Johnson, M., and Levinson S., "Automatic Detection of Contrast for Speech Understanding," *Proceedings of ICSLP 2004*, Jeju Island, South Korea, 2004.

[23]Liu, Y., Shriberg, E., and Stolcke, A., "Automatic Disfluency Identification in Conversational Speech Using Multiple Knowledge Sources," *Proceedings of Eurospeech*, Geneva, 2003.

[24]Beckman, M.E., Hirschberg, J., and Shattuck-Hufnagel, S. "The original ToBI system and the evolution of the ToBI framework," In: Sun-Ah Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*, Oxford University Press, Oxford, U.K., 2005.