Non-English Response Detection Method for Automated Proficiency Scoring System

Su-Youn Yoon and Derrick Higgins
Educational Testing Service
660 Rosedale Road, Princeton, NJ, USA
{syoon,dhiggins}@ets.org

Abstract

This paper presents a method for identifying non-English speech, with the aim of supporting an automated speech proficiency scoring system for non-native speakers.

The method uses a popular technique from the language identification domain, a single phone recognizer followed by multiple language-dependent language models. This method determines the language of a speech sample based on the phonotactic differences among languages.

The method is intended for use with non-native English speakers. Therefore, the method must be able to distinguish non-English responses from non-native speakers' English responses. This makes the task more challenging, as the frequent pronunciation errors of non-native speakers may weaken the phonetic and phonotactic distinction between English responses and non-English responses. In order to address this issue, the speaking rate measure was used to complement the language identification based features in the model.

The accuracy of the method was 98%, and there was 45% relative error reduction over a system based on the conventional language identification technique. The model using both feature sets furthermore demonstrated an improvement in accuracy for speakers at all English proficiency levels.

1 Introduction

We developed a non-English response identification method as a supplementary module for the automated speech proficiency scoring of non-native speakers. The method can identify speech samples of test takers who try to game the system by speaking in their native languages. For the items that elicited spontaneous speech, fluency features such as speaking rate have been one of the most important features in the automated scoring. By speaking in their native languages, speakers can generate fluent speech, and the automated proficiency scoring system may assign a high score. This problem has been rarely recognized, and none of research has focused on it as to the authors' knowledge. In order to address this issue, the automated proficiency scoring system in this study first filters out the responses in non-English languages, and for the remaining responses, it predicts the proficiency score using a scoring model.

Non-English detection is strongly related to language identification(Lamel and Gauvain, 1993; Zissman, 1996; Li et al., 2007); language identification is the process of determining which language a spoken response is in, while non-English detection makes a binary decision whether the spoken response is in English or not. Due to the strong similarity between the two tasks, the language identification method was used here for non-English response detection.

In contrast to previous research, the method described here was intended for use with non-native speakers, and the English responses for model training and evaluation were accordingly collected from non-native speakers. Among other differences, non-native speakers' speech tends to display non-standard pronunciation characteristics which can

Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, pages 161–169,
Portland, Oregon, 24 June 2011.

make the task of language identification more challenging. For instance, when native Korean speakers speak English, they may replace some English phonemes not in their language with their native phones, and epenthesize vowels within consonant clusters. Such processes tend to reduce the phonetic and phonotactic distinction between English and other languages. The frequency of these pronunciation errors is influenced by speakers' native language and proficiency level, with lower-proficiency speakers likely to exhibit the greatest degree of divergence from standard pronunciation. Language identification method may not effectively distinguish non-fluent speakers' English responses from non-English responses. In order to address these non-native speech characteristics, the model described here includes the speaking rate feature, which has been found to be an indicator of speaking proficiency in previous research(Strik and Cucchiarini, 1999; Zechner et al., 2009). Non-fluent speakers' English responses can be distinguished from non-English responses by slow speaking rate.

This paper will proceed as follows: we first review previous studies in section 2, then describe the data in section 3, and present the experiment in section 4. The results and discussion are presented in section 5, and the conclusions are presented in section 6.

2 Previous Work

Many previous studies in language identification focused on phonetic and phonotactic differences among languages. The frequencies of phones and phone sequences differ according to languages and some phone sequences occur only in certain languages. The literature in language identification captured this characteristic using the likelihood score of speech recognizers, which signals the degree of a match between the test sentences and speech recognizer models. Both the language model (hereafter, LM) and acoustic model (hereafter, AM) of a phone recognizer are optimized for the acoustic characteristics and the phoneme distribution of the training data. If a spoken response is recognized using a recognizer trained on a different language, it may result in a low likelihood score due to a mismatch between the test sentences and the models.

Lamel and Gauvain (1993) trained multiple language-dependent-phone-recognizers and selected the language with the highest matching score as the input language (hereafter, parallel PRLM). For instance, if the test data contained English and Hindi speech data, the English-phone-recognizer and the Hindi-phone-recognizer were trained independently. In the test, the given speech samples were recognized using two phone recognizers, and the language that had a higher matching score was selected. However, training multiple phone recognizers was time-consuming and labor intensive; therefore, Zissman (1996) proposed a system using single-language phone recognition followed by multiple language-dependent language modeling (hereafter, PRLM). PRLM was able to achieve comparable performance to parallel PRLM for long speech (longer than 30 seconds), and in a two-language situation, the error rate was between 5 and 7%.

Instead of language-dependent LM, Li et al. (2007) used vector space modeling (VSM). They applied metrics frequently used in information retrieval. As with the PRLM method, the speech was converted into phone sequences using the phone recognizer, and cooccurrence statistics such as term frequency (TF) and inverse document frequency (IDF) were calculated. The method outperformed the PRLM approach for long speech.

These methods can be challenging and time-consuming to implement, as they require implementation of methods beyond those typically available in a standard word-based recognition system. In particular, the application of the phone recognizer increases the processing time substantially. Because of this problem, Lim et al. (2004) presented a method based on the features that were readily available for speech recognizers: a confidence score and the cross-entropy of the LM. The confidence scoring method measured the acoustic match between the word hypotheses and the real sound, while the cross-entropy measured how well a sentence matched a given language model. If the test sentence was recognized by the speech recognizer in a different language, the phonetic and lexical mismatches between two languages resulted in a low confidence score and a high cross-entropy. Using this methodology, Lim et al. (2004) achieved 99.8% accuracy in their threeway classification task.

The current study can be distinguished from the previous studies in the following points. First of all, special features were implemented to model non-native speech since the method was developed for non-native speech. In our study, the data contained non-native speakers' English speech, characterized by inaccurate pronunciation. It resulted in a mismatch between the speech-recognizer models and test sentences, even for utterances in English. In particular, the mismatch was more salient in non-fluent speakers' speech, which comprised a high proportion of our data. In order to address this issue, speaking rate, which has achieved good performance in the estimation of non-native speakers' speaking proficiency (Strik and Cucchiarini, 1999; Zechner et al., 2009), was implemented as an additional feature. Secondly, in contrast to previous studies that determined which language the speech was in, we made a binary decision whether the speech was in English or not. Finally, the method was developed as part of a language assessment system.

3 Data

The OGI Multi-language corpus (Muthusamy et al., 1992), a standard language identification development data set, was used in the training and evaluation of the system. It contains a total of 1,957 calls from speakers of 10 different languages (English, Farsi, French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil, and Vietnamese). The corpus was composed of short speech and long speech; the short files contained approximately 10 seconds speech, while the long files contained speech ranged from 30 seconds to 50 seconds.

The method described here was implemented to distinguish non-English responses from non-native speakers' English responses. Therefore, the English data used to train and evaluate the model for non-English response detection was collected from non-native speakers. In particular, responses to the English Practice Test (EPT) were used. The EPT is an online practice test which allows students to gain familiarity with the format of a high-stakes test of English proficiency and receive immediate feedback on their test responses based on automated scoring methods. The speaking section of the EPT assessment consists of 6 items in which speakers are prompted to provide open-ended responses of 45-60 seconds in length. The scoring scale of each item is discrete from 1 to 4, where 4 indicates high speaking proficiency and 1 low proficiency.

The non-English detection task is composed of two major components: training of PRLM, and training of the classifier which makes a binary decision about whether a speech sample is in the English language, given PRLM-based features and the speaking rate.

The OGI corpus was used in training of both PRLM and the classifier; a total of 9,033 short files from the OGI corpus were used in PRLM training, and 158 long files were used in classifier training. (The small number of long files in the OGI corpus limited the number of samples comparable in length to our English-language data described below, so that only these 158 OGI samples could be used in classifier training and evaluation.) For English, only short samples were selected for use in this experiment.

In addition, a total of 3,021 EPT responses were used in classifier training. As the English proficiency levels of speakers may have an influence on the accuracy of non-English response detection, the EPT responses were selected to include similar numbers of responses for each score level. Responses were classified into four groups according to their proficiency scores and 1000 responses were randomly selected from each group. For score 1 and 4, where the number of available responses was smaller than 1000, all available responses were selected. Ultimately, 156 responses for score 1, 1000 responses for score 2 and score 3, and 865 responses for score 4 were selected.

The resultant training and evaluation data sets are summarized in Table 1.

Due to the lack of non-Engilsh responses in EPT data, 158 non-English utterances in OGI data were used in both training and evaluation of non-English detection. EPT responses were collected from many different countries, and speakers with 75 different native languages were participated in the data collection. Due to the large variations, many of their native languages were not covered by OGI data. However, all 9 languages in OGI data were in top 15 L1 languages and covered approximately 60% of speakers'

Partition name	Purpose	Number of English files	Number of non-English files
PRLM-train	Training of Language-dependent LM	1,716 (OGI)	7,317 (OGI)
EN-detection	Training and evaluation of non-English detection classifier	3,021 (EPT)	158 (OGI)

Table 1: Data partition

native languages.

4 Experiment

4.1 Overview

Due to the efficiency in processing time and implementation, a PRLM was implemented instead of a parallel PRLM. However, the difference between PRLM and parallel PRLM in this study may not be significant since PRLM has been shown to be comparable to parallel PRLM for test samples longer than 30 seconds, and the duration of test instances in this study was longer than 30 seconds. In addition to PRLM, speaking rate was calculated as a feature.

4.2 PRLM based features

The PRLM based method in this study is composed of three parts: training of a phone recognizer, training of language-dependent LMs, and generation of PRLM-based features. In contrast to the conventional language identification approach that only focused on identifying the language with the highest matching score, 6 additional features were implemented to capture the difference between English model and other models.

Phone recognizer: An English triphone acoustic model was trained on 30 hours of non-native English speech (EPT data) using the HTK toolkit (Young et al., 2002). The model contained 43 monophones and 4,887 triphones. Due to the difference in the sampling rate of EPT (11,025 Hz) and the OGI corpus (8,000 Hz), the EPT data was down-sampled to 8,000 Hz and the acoustic model was trained using the down-sampled data. In order to avoid the influence of English in phone hypothesis generation, a triphone bigram language model with a uniform probability distribution was used as the LM. (All possible combinations of two triphones were generated and a uniform probability was assigned to each combination.) The phone recognition accuracy rate was 42.7% on the 94 held-out EPT test samples. This phone recognizer was used in phone hypothesis generation for all data; the same recognizer was used for all languages.

Language-dependent LMs: For responses in the PRLM-train partition, phone hypothesis was generated using the English recognizer. Instead of the manual transcription, a language-dependent phone bigram LM was trained using the phone hypothesis. In order to avoid a data sparseness problem, triphones were converted into monophones by removing left and right context phones, and a bigram LM with closed vocabulary was trained. 10 language-dependent bigram LMs, including one for English, were trained.

PRLM based feature generation: For each response in the EN-detection partition, phone hypothesis was generated, and triphones were converted into monophones. Given monophone hypothesis, an LM score was calculated for each language using a language-dependent LM. A total of 10 LM scores were calculated.

Since the LM score increases as the number of phones increases, the LM score was normalized by the number of phones in each response, in order to avoid the influence of hypothesis length. 7 features were generated based on these normalized LM scores:

- MaxLanguage: The language with the maximum LM score

- SecondLanguage: The language with the second-largest LM score.

- MaxScore: Normalized LM score of the MaxLanguage.

- MaxDifference: Difference between normalized English LM score and MaxScore
- MaxRatio: Ratio between normalized English LM score and MaxScore
- AverageDifference: Difference between normalized English LM score and the average of normalized LM scores for languages other than English
- AverageRatio: Ratio between normalized English LM score and the average of normalized LM scores for languages other than English

Among above 6 features, 4 features (MaxDifference, MaxRatio, AverageDifference, and AverageRatio) were designed to measure the difference between matching of a test responses with English model and it with the other models. These features may be particularly effective when MaxLanguage of the English response is not English; these values will be close to 0 when the divergence due to non-native characteristics result in only slightly better match with other language than that with English.

4.3 Speaking rate calculation

The speaking rate was calculated as a feature relevant to establishing speakers' proficiency level, as established in previous research. Speaking rate was calculated from the phone hypothesis as the number of phones divided by the duration of responses (cf. Strik and Cucchiarini (1999)).

4.4 Model building

For each response, both PRLM-based features and speaking rate were calculated, and a decision tree model was trained to predict binary values (0 for English and 1 for non-English) using the J48 algorithm (WEKA implementation of C4.5) of the WEKA machine learning toolkit (Hall et al., 2009).

Due to the limited number of non-English responses in the EN-detection partition, three-fold cross validation was performed during classifier training and evaluation. The 3,179 responses were partitioned into three sets to include approximately same numbers of non-English responses and English responses for each proficiency score group. Each partition contained 52 $\sim$ 53 non-English responses and 1007 English responses. In each fold, the decision tree was trained using two of these partitions and tested on the remaining one.

5 Evaluation

First, the accuracy of the PRLM method was evaluated based on multiple forced-choice experiments with two alternatives using OGI data; in addition to non-English responses in EN-detection partition, English responses from the OGI data were used in this experiment. For each response (in English and one other language), phone hypothesis was generated and two normalized LM scores were calculated using the English LM and the LM for the other language. The MaxLanguage was hypothesized as the source language of the speech. The same experiment was performed for 9 combinations of English and other languages. Each experiment was comprised of 17 English utterances and 17 non-English utterances[1]. The majority class baseline was thus 0.5. The mean accuracy of the 9 experiments in this study was 0.943, which is comparable to (1996)'s performance: in his study, the best performing PRLM exhibited an average accuracy of 0.950. This initial evaluation used the same data and feature as Zissman (1996). (Of the seven PRLM-based features listed above, only MaxLanguage was used in (1996)'s study.)

Table 2 summarizes the evaluation results of the non-English response detection experiments using three-fold cross-validation within the EN-detection partition. In order to investigate the impact of different types of features, the features were classified into four sets—**MaxLanguage** only, **PRLM** (encompassing all PRLM features), **SpeakingRate**, and **all**—and models were trained using each set. The baseline using majority voting demonstrated an accuracy of 0.95 by classifying all responses as English responses.

All models achieved improvements over baseline. In particular, the model using all features achieved a 66% relative error reduction over the baseline of 0.95. Furthermore, the all-features model outperformed the model based only on PRLM or speaking

[1] Due to the languages where the available responses were only 17, the same 17 English responses were used in the all experiment although 18 responses were available

Features	Acc.	Pre.	Rec.	F-score
Baseline	0.950	0.000	0.000	0.000
MaxLanguage	0.969	0.943	0.411	0.572
PRLM	0.966	0.675	0.633	0.649
SpeakingRate	0.962	0.886	0.278	0.415
All	0.983	0.909	0.746	0.816

Table 2: Performance of non-English response detection

rate; the accuracy of the all-features model was approximately 1-2% higher than other models in absolute value and represented approximately a 45-50% relative error reduction over these models.

The PRLM-based model had higher overall accuracy than the speaking rate-based model, and the difference was even more salient by the F-score measure: the PRLM-based model achieved an F-score approximately 24% higher than the speaking rate-based model.

The model based on all PRLM features did not achieve a higher accuracy than the model based on only MaxLanguage. However, there was a clear improvement in F-score by using the additional features. The PRLM-based model achieved an F-score approximately 8% higher than the model based only on MaxLanguage.

In order to investigate the influence of speakers' proficiency on the accuracy of non-English detection, the responses in EN-detection were divided into 4 groups according to proficiency score, and the performance was calculated for each score group; the performance of each score group was calculated using subset comprised of all non-English responses and English responses with the corresponding scores.

A majority class baseline (classifying all responses as English) was again used. Table 3 summarizes the results observed, by score level, for the baseline model and for four different models used in Table 2. Note that the baseline is lower in Table 3 than in Table 2, because the ratio of English to non-English responses is lower for each of the subsets of the EN-detection partitions used for the evaluations at a given score level.

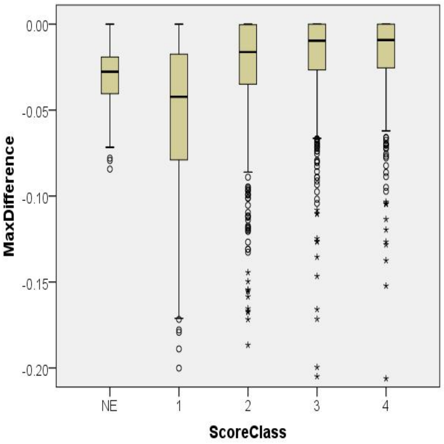

Figure 1: Relationship between proficiency score and MaxDifference

For all score groups, the model using all features achieved high accuracy. The model's accuracy on all data sets except for score group 1 was approximately 0.96 and the F-score approximately 0.85. The accuracy on score group 1 was 0.87, relatively lower than other score groups. This is largely due to the smaller number of English responses available at score level 1, and the consequent lower baseline on this data set. However, the relative error reduction was much larger; it was 74% for score group 1.

For all score groups, the PRLM-based models outperformed MaxLanguage based models and speaking rate based models. Additional PRLM features improved the performance over the models only based on MaxLanguage (conventional language identification method). In addition, the combination of both types of features resulted in further improvement.

The consistent improvement of the model using both PRLM and speaking rate features suggests a compensatory relationship between these features. In order to investigate this relationship in further detail, two representative features, MaxDifference and AverageDifference were selected, and boxplots were created. Figure 1 and Figure 2 show the relationship between proficiency score and PRLM features. In these figures, the label 'NE' is used to indicate the non-English group, while the labels 1, 2, 3, and 4 correspond to each score group.

Figure 1 shows that MaxDifference decreases as

Score	Features	Acc.	Pre.	Rec.	F-score
1	Baseline	0.497	0.000	0.000	0.000
	MaxLanguage	0.696	0.970	0.411	0.577
	PRLM	0.792	0.936	0.633	0.752
	SpeakingRate	0.636	1.000	0.278	0.432
	All	0.869	0.992	0.746	0.851
2	Baseline	0.865	0.000	0.000	0.000
	MaxLanguage	0.919	0.983	0.411	0.579
	PRLM	0.930	0.811	0.633	0.709
	SpeakingRate	0.901	1.000	0.278	0.432
	All	0.962	0.971	0.746	0.843
3	Baseline	0.865	0.000	0.000	0.000
	MaxLanguage	0.920	1.000	0.411	0.582
	PRLM	0.939	0.903	0.633	0.738
	SpeakingRate	0.901	0.983	0.278	0.430
	All	0.963	0.976	0.746	0.845
4	Baseline	0.846	0.000	0.000	0.000
	MaxLanguage	0.908	0.987	0.411	0.579
	PRLM	0.936	0.934	0.633	0.752
	SpeakingRate	0.882	0.896	0.278	0.417
	All	0.955	0.956	0.746	0.837

Table 3: Performance of non-English detection according to speakers' proficiency level

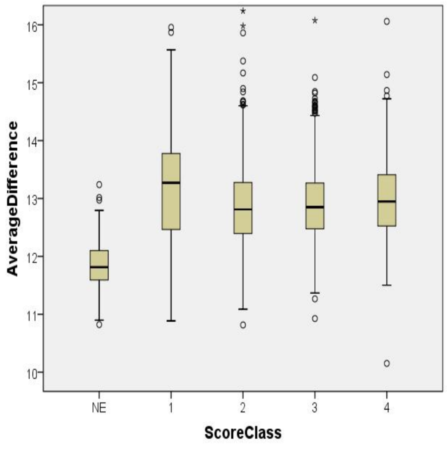

Figure 2: Relationship between proficiency score and AverageDifference

the speaker's proficiency decreases, although the feature displays a large variance. The feature mean for non-English responses is lower than for score groups 2, 3, and 4, but the distinction between non-English and English becomes smaller as the proficiency score decreases. The feature mean for score group 1 is even lower than for non-English responses. This obscures the distinction between English responses and non-English responses at lower score levels.

As Figure 2 shows, AverageDifference is relatively stable across score levels, compared to MaxDifference. Although the mean feature value decreases as the proficiency score decreases, the decrease is smaller than for MaxDifference. In addition, the mean feature values of the English groups are consistently higher than those for non-English responses.

Figure 3 shows the relationship between proficiency score and speaking rate.

For the speaking rate feature, the distinction between non-English and English responses increases as speakers' proficiency level decreases, as shown

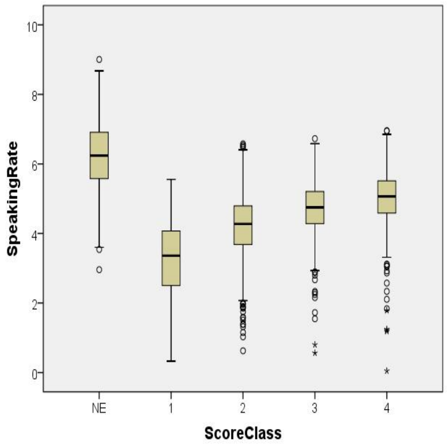

Figure 3: Relationship between proficiency score and SpeakingRate

in Figure 3. The speaking rate of non-English responses is the highest among all groups compared, and the speaking rate decreases for English responses as the speaker's proficiency score decreases. Thus, the PRLM features tend to display better discrimination between English and non-English responses at the higher end of the proficiency scale, while the SpeakingRate feature provides better discrimination at the lower end of the scale. By combining both feature classes, we are able to produce a model which outperforms both a PRLM-based model and a model using speaking rate alone.

6 Conclusion

In this study, we presented a non-English response detection method for non-native speakers' speech. A decision tree model was trained using PRLM-based features and speaking rate.

The method was intended for use as a supplementary module of an automated speech proficiency scoring system. The characteristics of non-native English speech (frequent pronunciation errors) reduced the phonetic distinction between English responses and non-English responses, and correspondingly, the differences between the feature values for non-English and English speech decreased as well.

In order to address this issue, a speaking rate feature was added to the model. This feature was specialized for second language (L2) learners' speech, as speaking rate has previously proved useful in estimating non-native speakers' speaking proficiency. In contrast to PRLM-based features, the speaking rate feature showed increasing discrimination between non-English and English speech samples as speakers' proficiency level decreased. The complementary relationship between PRLM-based features and speaking rate led to an improvement in the model when these features were combined. Improvements resulting from the combined feature set extended across speakers at all proficiency levels studied in the context of this paper.

The speaking rate becomes less effective if test takers speak slowly in their native languages. However, the test takers are unlikely to use this strategy, since it will result in a low score although they can game the system.

Due to lack of non-English responses in EPT data, non-English utterances were extracted from OGI data. Since the features in this study were not directly related to acoustic scores, the acoustic differences between EPT and OGI data may not give significant impact on the results. However, in order to avoid any influence by differences between corpora, the non-English responses will be collected using EPT setup and the evaluation will be performed using new data in future.

References

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. In *SIGKDD Explorations*, volume 11.

Lori F. Lamel and Jean-Luc Gauvain. 1993. Cross-lingual experiments with phone recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 507–510.

Haizhou Li, Bin Ma, and Chin-Hui Lee. 2007. A vector space modeling approach to spoken language identification. *Audio, Speech and Language Processing*, 15:271 – 284.

Boon Pang Lim, Haizhou Li, and Yu Chen. 2004. Language identification through large vocabulary continuous speech recognition. In *Proceedings of the 2004 International Symposium on Chinese Spoken Language Processing*, pages 49 – 52.

Yeshwant K. Muthusamy, Ronald A. Cole, and Beatrice T. Oshika. 1992. The OGI multi-language telephone speech corpus. In *Proceedings of the Inter-*

national Conference on Spoken Language Processing, pages 895–898.

Helmer Strik and Catia Cucchiarini. 1999. Automatic assessment of second language learners' fluency. In *Proceedings of the 14th International Congress of Phonetic Sciences*, pages 759–762, San Francisco, USA.

Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. 2002. *The HTK Book (for HTK Version3.2)*. Microsoft Corporation and Cambridge University Engineering Department.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(10):883 – 895.

Marc A. Zissman. 1996. Comparison of four approaches to automatic language identification of telephone speech. *Speech and Audio Processing*, 4:31 – 44.