A SEGMENTAL APPROACH TO AUTOMATIC LANGUAGE IDENTIFICATION

Yeshwant Kumar Muthusamy
B.Tech., Jawaharlal Nehru Technological University, Hyderabad, India, 1987

A dissertation submitted to the faculty of the
Oregon Graduate Institute of Science & Technology
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Computer Science and Engineering

October 1993

The dissertation "A Segmental Approach to Automatic Language Identification" by Yeshwant Kumar Muthusamy has been examined and approved by the following Examination Committee:

Ronald A. Cole, Professor
Research Adviser

Mark A. Fanty, Research Assistant Professor

Todd K. Leen, Associate Professor

Jonathan Walpole, Assistant Professor

Beatrice T. Oshika, Professor
Portland State University

Dedication

To my parents, Viveka and Ponnusamy Muthusamy.

Acknowledgements

First and foremost, I would like to thank Ronald Cole, my research adviser and mentor for the past six years. I am grateful to him for introducing me to the area of speech recognition, for teaching me everything I know about doing research and for treating me more like a friend than a student. Working with him has been a truly memorable experience.

I would also like to thank the other members of my committee: Mark Fanty, Todd Leen, Beatrice Oshika and Jonathan Walpole. Mark helped give shape and form to several of my nascent research ideas. I am grateful to him for not only patiently answering innumerable questions, but for also imparting some of his programming expertise to me. His insightful comments and insistence on cogency and clarity have enriched this dissertation tremendously. Todd added a rigorous mathematical dimension to my informal knowledge of neural networks and clarified my doubts on principal component analysis. I shall forever remain fascinated by Beatrice's expertise in linguistics and phonology. She always made time to share her knowledge of languages with me and refine the linguistic aspects of this research. Without her help, the ten-language telephone speech corpus would not have become a reality. Jonathan's abhorrence of excessive jargon helped to make this dissertation more readable. Special thanks to Etienne Barnard, the author of "opt", the neural network simulation package that I used heavily in my research. I am also grateful to him for pointing out loose ends in my research and for refining my dissertation with his comments.

Thanks also to Kay Berkling, Daniel Burnett, Terri Durham, Vincent Weatherill, Taimi Ames, Nandakishore Kambhatla, Johan Schalkwyk and other members of the OGI Center for Spoken Language Understanding for their help and friendship. I am

grateful to them for their tolerance of my copious use of disk space and CPU time. Terri's help with the conduct of the human listening experiments is much appreciated.

I thank Jon Inouye for selflessly expending time and effort to create the LaTeX versions of the OGI thesis style files. His hard work made the typesetting of this dissertation an effortless task. I also thank him for his genuine friendship and help on numerous occasions.

I owe a debt of gratitude to past and present OGI Computer Science and Engineering system and administrative staff: Nike Horton, Bruce Jerrick, Marion Hakanson, Nora Auseklis, John Pochmara, Marie LaBonte, Kelly Atkinson and Jo Ann Binkerd. Bruce, Kelly and Nike have not only been patient with my incessant stream of questions and special requests, but have also extended me their friendship.

I shall remain indebted to my parents for their guidance and support. Though thousands of miles apart for the most part of these last six years, I have always felt their presence by my side, in all my endeavors. This work would not have been possible without their affection and constant encouragement. My brother Karthik's impish sense of humor served to relieve the tedium of graduate student life. I am grateful to my wife, Sujaritha, for her love and patience, and most of all, for believing in me.

Finally, a belated word of thanks to my grandfather, C.V. Dakshina Murthy and my grand uncle, C.V. Kannan, for instilling in me the desire to emulate their academic achievements. I know they would have been proud of what I have accomplished.

Contents

List of Tables

List of Figures

Abstract

A SEGMENTAL APPROACH TO AUTOMATIC LANGUAGE IDENTIFICATION

Yeshwant Kumar Muthusamy, Ph.D.
Oregon Graduate Institute of Science & Technology, 1993

Supervising Professor: Ronald A. Cole, Professor

Automatic language identification is the problem of identifying the language being spoken from a sample of speech by an unknown speaker. A segmental approach to automatic language identification is based on the assumption that the acoustic structure of languages can be estimated by segmenting speech into phonetic categories. Language identification can then be achieved by computing features within and across segments that describe the phonetic and prosodic characteristics of individual languages, and using these feature measurements to train a classifier to distinguish between the languages. Recognizing the difficulties involved in the development of a phonetically labeled corpus of speech, we have applied this approach using broad phonetic categories.

This dissertation addresses the following questions: What acoustic, broad phonetic and prosodic information is needed to achieve automatic identification of languages? What is the best way to present this information to neural network classifiers? What is the level of language identification possible given only this information?

In preliminary research, this broad phonetic approach was applied to a four-language (English, Japanese, Mandarin and Tamil) corpus of high quality speech. The results of

this research were sufficiently promising to merit further investigation of the approach with a ten-language corpus of telephone speech consisting of mostly fluent speech from 90 speakers each of English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese.

Several features based on pairs and triples of broad phonetic categories were evaluated. Pitch-based features were found to perform the worst, while features based on pairs of broad phonetic categories performed the best.

Perceptual experiments were also conducted, in which trained listeners identified excerpts of speech of one-, two-, four-, and six-second durations as one of the ten languages. The results revealed that for some languages like Korean, Farsi and Vietnamese, identification performance was poor regardless of the duration of the excerpts.

The automatic identification results indicate that while broad phonetic categories do possess language discriminatory information, the level of identification performance possible with broad phonetic information alone leaves much to be desired. Information at the phonemic or phonetic level might be required to distinguish between languages with greater accuracy.

Chapter 1

Introduction

1.1 The Problem

Automatic language identification is the problem of identifying the language being spoken from a sample of speech by an unknown speaker. As with speech recognition, humans are the most accurate language identification systems in the world today. Within seconds of hearing speech, people are able to determine whether it is a language they know. If it is a language with which they are not familiar, they often can make subjective judgments as to its similarity to a language they know, e.g., "sounds like German". This determination is made so quickly that it is likely that phonological characteristics, as opposed to word or phrase recognition, are the basis for the decision.

It is well known that languages have characteristic sound patterns. Languages have been described subjectively as "singsong", "rhythmic", "guttural", "nasal" etc. Languages differ in the inventory of phonological units used to produce words, in their frequency of occurrence, and the order in which they occur in words. The presence of individual sounds, such as the "clicks" found in some sub-Saharan African languages, or the velar fricatives found in Arabic, are readily apparent to speakers of languages that do not contain these phonemes. Less obvious acoustic patterns are also observed. Mandarin Chinese has a higher frequency of occurrence of nasals than English. Hawaiian is known for its very limited consonant inventory. Prosodic patterns also differ significantly between languages. For example, it has been shown that fundamental frequency (F_0) patterns of continuous speech display different characteristics in Mandarin Chinese

(a tone language) and American English (a stress language) [Ead82]. The key to solving the problem of automatic language identification then, is the detection and exploitation of such differences between languages.

1.2 Motivation

The Workshop on Spoken Language Understanding sponsored by the National Science Foundation [C+92] has identified multi-lingual systems as one of the key research areas in the field of spoken language systems. Stressing the need to move towards multi-lingual speech research in an increasingly global economy, the workshop report lists automatic language identification as one of the challenges:

> As an independent capability or as part of a multi-lingual spoken language system, techniques are needed to identify language and/or dialect in order to route the user to the appropriate human (e.g., human telecommunications operator) or automatic system (e.g., spoken language data retrieval system).

What makes this problem so challenging and interesting? In mono-lingual spoken language systems, the objective is to determine the content of the speech, i.e., phoneme recognition followed by word recognition, followed by sentence recognition. This requires that researchers cue in on small portions of the speech—frames, phonemes, syllables, sub-word units, and so on, to determine what the speaker said. In contrast, in text-independent language identification, phonemes and other sub-word units alone are not sufficient cues, since several phonemes and syllables and even words are common across different languages. One also needs to examine the sentence as a whole to determine the "acoustic signature" of the language, the unique characteristics that make one language sound distinct from another. The fact that humans are so adept at this task illustrates the considerable gap between our perceptual capabilities and our attempts at automating them.

Aside from the fact that it is a challenging area of research, there are several important applications for automatic language identification. Much of the past funding for

research in this area has been provided by government agencies interested in communications monitoring for national security purposes. However, there are also important commercial demands. As the global economic community expands, there is an increasing need for automatic spoken language identification services. For example, checking into a hotel, arranging a meeting or making travel arrangements can be difficult for non-native speakers. Telephone companies will be better equipped to handle foreign language calls if an automatic language identification system can be used to route the call to an operator fluent in that language.

Rapid language identification and translation can even save lives. There are many reported cases of 911 operators being unable to understand the language of a distressed caller. In response to these needs, AT&T recently introduced its *Language Line* Interpreter Service to serve business, the general public and police departments handling 911 emergencies. The service uses trained human interpreters, handles 140 languages and satisfies an important need in our increasingly cosmopolitan communities. However, tremendous responsibility is placed on the human operator who must route the call to the appropriate interpreter. A call to the *Language Line* Service by the author, who spoke only in Tamil, resulted in a 3 minute delay before the language was identified and a Tamil interpreter was brought on-line. The delay was caused by the operator unsuccessfully trying out three South-East Asian interpreters and playing recordings of greetings in other languages. The delay would have been longer if the author had not relented and spoken the name 'Tamil' in English rather than in Tamil! This anecdote emphasizes the point that if automatic language identification could be made sufficiently fast and accurate, it could aid human operators.

An automatic language identification system could also serve as a front-end for a multi-language translation system in which the input speech can be in one of several languages. The input language needs to be quickly identified before translation to the target language(s) can begin.

1.3 A Segmental Approach

The workshop report goes on to identify several fundamental scientific issues that need to be addressed to meet the challenge of multi-lingual speech systems. Two issues relevant to language identification are worth quoting:

> 1. The general question of what are the fundamental acoustic, perceptual, and linguistic differences among languages that should be investigated, with a view toward accommodating these differences in multi-lingual systems.
>
> 2. An investigation should be undertaken of language-specific versus language-independent properties across languages. ...

It is safe to say that any approach to solving the language identification problem should address both of the issues outlined above to be assured of any success at all. In the next few paragraphs, I describe a segmental approach that examines differences between languages at the broad phonetic and prosodic level.

A segmental approach to automatic language identification assumes that each language has a unique acoustic signature, and that this signature can be defined in terms of segmental and prosodic features of speech. Segmental features include the inventory of broad phonetic segments such as vowels, fricatives, stops, nasals and closures, and their frequency of occurrence and co-occurrence in speech. Prosodic information consists of the relative durations and amplitudes of sonorant (vowel-like) segments, their spacing in time, and patterns of pitch change within and across these segments.

To the extent that these assumptions are valid, languages can be identified automatically by segmenting speech into broad phonetic categories, computing segment-based features that capture the relevant phonetic and prosodic structure, and training a classifier to associate the feature measurements with the spoken language. The segment-based features need to be linguistically motivated, i.e., based on an analysis of the language-specific and language-universal properties of the broad phonetic sequences. In addition, the use of segments as anchors for prosodic analysis of speech should augment the power of this approach.

One might argue that a fine phonetic approach to automatic language identification is likely to yield better identification accuracy. As indicated earlier, informal evidence suggests that human listeners cue in on specific sound units or phonemes unique to a language in making their recognition decision. However, a fine phonetic approach to text-independent automatic language identification has two major problems. First, accurate phoneme recognition from continuous speech is a difficult task, due to coarticulation effects that result in widely different articulatory and acoustic realizations of the same phoneme. The current state-of-the-art in speaker-independent phonetic recognition of American English is 78.0% for 39 phones [CL92] and 70.0% when all 61 TIMIT [LKS86, FDGM86] phones are considered [RF91]. Secondly, since languages differ in their phonetic inventories, a multi-language fine phonetic recognizer is essential to process speech from more than one language. Such a recognizer would have to be trained on a multi-language phonetic alphabet. While the International Phonetic Alphabet (IPA) [Ass89] might serve our purpose, human transcriptions or correction of the fine phonetic labels produced by automatic labeling algorithms are error-prone and time-consuming tasks, beyond the reach of our current resources.

On the other hand, it is clear that broad phonetic category information can be utilized to distinguish between languages [HN77, LE80], allowing us to circumvent the cumbersome and difficult task of fine phonetic recognition. An additional advantage is that broad phonetic categories such as vowels, obstruents and nasals are relatively invariant across languages, eliminating the need for language-specific phonetic or phonemic alphabets. Three questions then arise: what are the kinds of acoustic, broad phonetic and prosodic information that are needed to achieve automatic identification of languages? What is the best way to present this information to neural network classifiers? What is the level of language identification accuracy possible given only this information? This dissertation addresses just these questions.

1.4 Dissertation Overview

Chapter 2 surveys the surprisingly small amount of literature in automatic language identification and examines the reasons for the lack of published research. Chapter 3 describes the in-house collection and development of two multi-language speech corpuses that enabled the research for this dissertation. Chapter 4 describes a feasibility study: the development of a four-language automatic language identification system using high-quality (laboratory microphone) speech. The signal representations used, neural network architecture and details of two broad phonetic feature-based approaches are presented. Chapter 5 describes research on a ten-language automatic language identification system using speech recorded over commercial telephone lines. Experiments with broad phonetic, spectral and pitch features merit separate sections, each with its own discussion of the results of that feature set. The chapter ends with an analysis of the classifier performance on the combined feature set. Chapter 6 summarizes the findings of this dissertation and outlines directions for future research.

Chapter 2

Literature Review

While the past two decades have witnessed tremendous advances in automatic speech recognition with thousands of published papers, there have been only fourteen published papers in English in automatic language identification that the author is aware of. This chapter provides brief descriptions of these studies and critiques them. The last section points out the problems associated with the general trend of research in this area and proposes a two-pronged research strategy.

2.1 Previous Work

This literature review is confined to studies that were done before the start of the research reported in this dissertation. As a result, the review does not contain any mention of the author's published work on automatic language identification using high-quality speech [MC92a] or telephone speech [MC92b]. These are described in chapters 4 and 5 respectively. Marc Zissman's work [Zis93] and Hazen and Zue's work [HZ93] using the OGI Multi-language Telephone Speech Corpus ([MCO92] and Chapter 3) are described in Chapter 6. Also, since the focus of this review is on studies in *automatic* language identification, Atkinson's work on human listening experiments with English and Spanish [Atk68] is excluded.

The studies were reviewed keeping in mind the following variables of importance:

- number and confusability of languages: the fewer the number of languages to identify, the easier the problem

- content independence: whether the vocabulary of the speech was fixed or unrestricted
- speaker-independence: whether there was any overlap between the training and test speakers
- speech corpuses used, including recording conditions, type of speech (read or spontaneous), sampling rate, signal-to-noise ratio, and number and background of speakers (i.e., native or non-native)
- primary methodology: HMMs, neural networks, expert systems, clustering, etc.
- granularity of feature measurement: global features or a moving analysis window?
- amount of speech needed for the stated identification accuracy
- gender-specificity of the results: whether the algorithms were trained and evaluated on both male and female speakers

Table 2.1 summarizes the salient features of the studies. The reader is directed to the references for further details.

2.1.1 The Texas Instruments Effort

The first sustained effort in automatic language identification was carried out between 1973 and 1980 at Texas Instruments (TI), and is documented in a series of four reports [LD74, LD75, LD78, Leo80].

The basic philosophy underlying the TI approach was that languages differ by the frequency of occurrence of certain reference sounds or sound sequences. The sounds or sound sequences characteristic of a language occur more often in that language than in any other language under consideration. Therefore, the likelihoods of the languages, given these sequences, could be computed and used to make decisions in reasonably short times.

Table 2.1: Studies in Automatic Language Identification

STUDY	LANGUAGES	TYPE OF DATA	SPEAKERS	APPROACH	RESULTS
Texas Instruments (1973-80)	7 (not specified)	Read Speech	100 adult males (50 train 50 test)	Detection of "Reference Sounds" and estimation of log likelihoods of the languages	62% (no rejection) 100% (68% rejection)
House and Neuberg (1977)	American English, Chinese, Greek, Korean, Urdu, Japanese, Russian and Swahili (8)	Phonetic trans. of text (no real speech)	-	HMMs trained on sequences of broad category labels	Near-perfect discrimination (no % specified)
LI and Edwards (1980)	2 Asian & 3 Indo-European (not specified)	Read Speech	150 (50 train 50 evaluate 50 test)	Segment-based and syllable-based Markov models	80%
Cimarusti and Ives (1982)	American English, Czech, Farsi, Korean, German, Mandarin, Russian & Vietnamese (8)	Read Speech	40 (train and test sets unspecified)	Acoustic features and a polynomial decision function	84%
Ives (1986)	American English, Czech, Farsi, Korean, German, Mandarin, Russian & Vietnamese (8)	Spoken Speech (100 to 5000 Hz)	122 adult males (train and test sets unspecified)	Expert System Production Rules	92%
Foil (1986)	3 (not specified)	Speech from radio (SNR 5 dB)	Not specified	Processing Pitch & Energy Contours Formant-clustering algorithm	39% 64 % (11% rejection)
Goodman et. al (1989)	Four different sets of languages (not specified)	Speech from radio (SNR 9 dB)	Not specified	Improved Formant Clustering Algorithm	Reduced Foil's error rate in half (no % specified)
Sugiyama (1991)	20 languages (CCITT SG-XII CD-ROM)	Spoken speech (Avg. SNR 49.2 dB)	76 Males 77 Females	Standard VQ VQ histogram Algorithm	65% 80%
Savic et. al (1991)	English, Hindi, Mandarin & Spanish	Read Speech (0 - 4.5 kHz)	Not specified	HMMs & Pitch Contour Analysis	Not specified
Nakagawa et. al (1992)	English, Japanese, Mandarin & Indonesian	Conversational Speech (0 - 6 kHz)	60 Males	HMM-based methods	86.3%
Kwasny et. al (1992)	English & French	Read Speech (0 - 12 kHz)	1 Male 1 Female	Acoustic Features and ANNs	100%

Study 1. The first study [LD74] concentrated on single reference sounds. The data consisted of read text from 100 adult male speakers of 5 languages, referred to simply as L_1, L_2, L_3, L_4, and L_5. The training data consisted of 90-second segments of speech from each of 10 speakers of each of the five languages. The test data consisted of 90-second segments from: 10 speakers of L_1, L_3, and L_5; 6 speakers of L_2; and 14 speakers of L_4. The recording conditions and sampling frequency were not specified.

The first step in this approach was automatic segmentation of the digitized speech based on a measure of dynamic spectral change T called "transitionitivity". Transitionitivity T was defined as the squared error between two sound segments displaced from each other by 3 time intervals, with the displacement being determined experimentally.

Reference files of sound segments potentially useful for language discrimination were automatically generated from the training data, using an "intersegment distance" measure. This technique allowed a segment to be added to a file only if it was sufficiently different from each segment already in the file. These reference files were then "purged", either to actually eliminate infrequently occurring segments, or to retain the segments in the file, but to use them only for partitioning the data. One form of purging was to set a predetermined threshold for the number of occurrences of each segment, and delete segments that fell below this threshold.

The frequency of occurrence of the remaining reference segments in the files was determined and the time-averaged log-likelihood of the languages was computed (i.e., for each language L and reference segment R, the computation yielded the likelihood that language L was spoken, given that segment R had occurred). In one experiment, decision functions were computed for each pair of languages for each of the 50 test speakers. The decision strategy was to choose the language with the smallest negative average log-likelihood. Pairwise identification accuracy of the 10 language pairs ranged from 60% to 100%. Overall accuracy was 64% with a nearest neighbor decision rule using the pairwise identification results. The identification decision was made using 60 seconds of speech.

Study 2. The second phase of the study [LD75] used the same data as above, but used *sequences* of several phoneme-like segments for classification. Another improvement was the use of a time-frequency scanning error measure to accept or reject hypothesized occurrences of component sound segments. The scanning error was a measure of spectral similarity between the input speech data and scanning patterns representing each candidate reference segment. Two measures were introduced to help prune the file of reference sequences: (i) an information-theoretic measure called "entropy threshold" guided the selection of reference sequences with sufficient language specificity, and (ii) an acceptance level for hypothesized sequences, that rejected sequences that did not occur often enough to merit inclusion. The entropy threshold was used to determine the language specificity of sequences by considering the average information remaining (uncertainty, entropy) after the detection of a reference sequence in the training data. The lower this uncertainty, the better the language discrimination capability of that sequence.

Classification was based on the summed logarithms of the language likelihood estimates, given the occurrences of the reference sound sequences.

Experiments were performed with sound sequences of different lengths. It was found that sequences of length 4 performed best on the training data: 88% correct classification of the 5 languages with an entropy threshold of 2.3 and acceptance level of 12.5%. A decision rule using sequences of length 5 in combination with sequences of length 1 yielded 70% accuracy on the test data, with the same threshold values as above.

Study 3. Departing from the automatic selection of reference sounds used in the first two studies, the third study [LD78] used an interactive approach to the generation of reference sounds.

Manual selection of reference sounds was followed by automatic isolation of the representative occurrences of these sounds from the speech data. The isolated sounds were then manually verified before further processing. Using 36 reference sounds and a silence measure, a nearest-mean decision rule was applied to yield 94% correct classification on the training set. Various subsets of these 37 features were evaluated on the test set to

determine the subset that provided maximum performance. A subset of 13 reference sounds and the silence measure yielded 66% on the 50 test set speakers.

The data collection documentation indicated that five different sources of data and three different sets of recording equipment were used to collect speech from the 100 speakers used in the experiments. In an attempt to reduce variations due to different recording conditions, each speaker's long-term average spectrum was standardized. This resulted in much improved identification performance. A second set of 13 reference sounds and the same silence measure provided 80% correct classification on the test set speakers.

Study 4. In the final study [Leo80], the interactive approach to reference sound generation was extended to allow more accuracy in specifying reference sounds and more flexibility in the allowed types of reference sounds. Another improvement was the introduction of a criterion for rejection, i.e. not classifying an utterance when the basis for such a decision was not sufficiently strong.

The data of the first three studies were augmented with speech from 17 speakers of language L_7 (L_6 was reserved for English) and 14 speakers of language L_8. There were now 66 speakers in the training set and 65 speakers in the test set.

The speech was digitized and the characteristic sound sequences determined using the improved interactive reference sound generation program. The initial reference file had 94 sounds from the 7 languages. The training data was processed to automatically detect and count occurrences of these sounds to compute parameters of a decision function. After applying pruning techniques based on various thresholds, a file of 80 reference sounds was produced. The decision function for a language was defined to be the negative of the sum of the log-likelihoods for all detected reference sounds. The test data was then processed to detect and count reference sounds to evaluate decision function values (one for each possible language). The language with the minimum decision function value was chosen. If the difference between the smallest and next smallest decision function value was below a certain threshold, the speaker was rejected (the rejection criterion).

With the 7-language 65-speaker test set consisting of 80 reference sounds, 62% accuracy was attained when no rejects were allowed, and 100% accuracy was achieved with a rejection rate of 68%. With the original 5-language 50-speaker test set consisting of 54 reference sounds, the corresponding figures were 72%, and 100% with a rejection rate of 56%.

Critique. It is clear that significant contributions to the field of automatic language identification have been made by the TI effort. The TI studies embody notions of phonetic distinctiveness of languages. They followed a "phoneme spotting" approach using automatically defined sub-word units and measures to indicate which units were discriminative. Seven years of sustained effort produced improvement from 64% to 80% on the five-language task.

The classified nature of the studies raises more questions than it answers. What were the seven languages? What were the reference sounds? Without this information, it is difficult to evaluate the results of this work. Moreover, the extensibility of their general approach is open to question. Improved results were obtained in the latter two studies, in which automatic determination of reference sounds was replaced by an interactive process that required considerable human input. However, such manual determination of the reference sounds in the languages under consideration mandates the researchers' *a priori* knowledge of the languages. This could severely limit addition of languages to the identification system. This weakness is apparent in the fourth study [Leo80] in which there is a degradation in performance (from 72% to 62%) with the addition of the two languages L_7 and L_8. The author attributes it to a lack of *familiarity* with the two languages resulting in selection of inappropriate reference sounds for these languages.

2.1.2 House and Neuberg

In a landmark study, House and Neuberg [HN77] demonstrated the feasibility of using sequences of broad phonetic categories of speech to identify languages. They reasoned that, since accurate phoneme recognition is beyond the current state-of-the-art (true

in 1977; true in 1993), the information provided by broad phonetic categories (stop, fricative, vowel, silence) should be examined. They assumed that the sequence of broad phonetic categories of a language could be modeled by a Markov process, and that the parameters of the model could be estimated for a given language from sufficient training data.

The data for this study consisted of manually generated phonetic transcriptions of text from each of the following eight languages: American English, Chinese, Greek, Japanese, Korean, Russian, Swahili and Urdu.

Statistical models were trained on sequences of broad phonetic category labels derived from these phonetic transcriptions. Given a sequence of broad category symbols U, the probability $P(U|L)$ is calculated for each language L, and U is said to represent the language L for which this probability is a maximum. The texts were used in their entirety to train the models, leaving no text against which the model for a given language could be tested. To overcome this problem, each text was tested against every model (including its own) in an attempt to establish the power of this procedure to distinguish among languages. Perfect discrimination of the eight languages was obtained.

Critique. House and Neuberg's study is an important concept demonstration, and shares many of the assumptions of the approach examined in this dissertation. Their actual study, however, did not make use of real speech, only phonetic transcriptions of *text*. The study assumes perfect segmentation and classification of broad phonetic segments which is not possible with real speech data. Another caveat is the lack of an independent test set against which the language models could be tested. Further, even though perfect discrimination between the languages was achieved, the differences between the individual language scores were very small, raising questions about their statistical significance.

2.1.3 Li and Edwards

The Markov techniques suggested by House and Neuberg were further developed by Li and Edwards and applied to real speech data. Their work [LE80] represents one of the earliest efforts to develop statistical inference techniques to discriminate among languages using real speech data . They used a broad segmentation scheme to classify data into six acoustic-phonetic classes: (i) syllabic nuclei, (ii) non-vowel sonorants, (iii) vocal murmur, (iv) voiced frication, (v) voiceless frication, and (vi) silence and low energy segments.

Based on these broad segmental classes, two statistical models for automatic language identification were developed: one based on segments and one based on syllables. The segmental models were implemented as either zero, first or second order Markov models that characterized segmental sequences in the languages.

The syllable model was divided into two types, one based on inter-syllable-nuclei sequences and one based on intra-syllable-nucleus segment sequences. The inter-syllable zero-order Markov model described segment sequences between two syllabic nuclei, which can be roughly paraphrased as characterizing possible consonant clusters in the languages. The intra-syllable model represented a syllable as a nucleus preceded or followed by up to two segments (not including a neighboring syllabic nucleus), and approximated the internal structure of a syllable without requiring detection of specific syllable boundaries. The intra-syllable model was implemented as both zero and first order Markov models.

The database consisted of read speech from 20 speakers of five languages, two Asian and three Indo-European. The two Asian languages were basically monosyllabic tonal languages with relatively simple consonant-vowel (CV) or CVC word structure. The three European languages represented two different language families, and were distinguished from the Asian languages by greater word length and more complex consonant clusters.

The training database consisted of 200 minutes of speech (four minutes each from ten speakers for each of five languages) collected in a reading mode, for a total of about

42,000 syllables and 150,000 segments. The test data was 100 minutes of read speech (two minutes each from ten speakers for five languages). All speakers were male. The recording conditions and sampling frequency were not specified.

The identification procedure consisted of moving a variable length analysis window through the training data and the independent test data. The analysis window was x "segments" (for the segment-based model) or y "syllables" (for the syllable-based models) where x and y were varied to cover an analysis period from 15 seconds to two minutes long. Each model was tested over a selected analysis window with each language accumulating a conditional probability of being the language tested. For each window, an accumulated weighted vote was obtained for each language based on the conditional probabilities. The window was then incremented through the test data by one element (segment or syllable) and the process repeated with new weighted votes accumulated until the data was exhausted for each speaker. The language associated with the largest analysis-window vote for that speaker was chosen as the correct language.

The results of these techniques varied considerably across the various models, reaching a maximum of about 80% correct identification using the inter-syllable model for an independent test of 50 speakers (10 per language). An analysis of the confusions among languages indicated that the techniques distinguished the two major types of languages very well, that is, the Asian languages from the Indo-European languages. This suggests that a two-stage algorithm might be useful in language identification. The first stage divides the languages into major types, and the second stage examines the languages within each type in more detail and makes focused decisions based on known characteristics of that language type.

Critique. This study extended House and Neuberg's concept demonstration, by applying it to real speech data. The classification accuracy of 80% on five languages validates House and Neuberg's hypothesis that broad phonetic category sequences do possess language discriminatory information. On the other hand, the results of this study cannot be

meaningfully compared with those of other studies, since the languages used, the recording conditions and the sampling frequency of the digitized speech are all unknown. Also, the lack of female speakers in the database and the use of read speech limit the generality of the results.

2.1.4 Cimarusti and Ives

Cimarusti and Ives [CI82] conducted a feasibility study of a novel approach to automatic language identification that was not based on linguistic units such as phonetic segments or syllables. This approach applied pattern analysis techniques to acoustic features extracted from the speech signal.

The data consisted of three minutes of read speech collected from audio recordings of five adult male speakers for each of the following eight languages: American English, Czech, Farsi, German, Korean, Mandarin, Russian and Vietnamese. The audio recordings were digitized at a sampling frequency of 10 kHz and passed through a 5 kHz anti-aliasing filter. The data was randomly divided into training and test sets.

Using a 30 ms moving analysis frame with a 30 ms increment, 100 features derived from LPC analysis (including autocorrelation coefficients, cepstral coefficients, filter coefficients, log area ratios and formant frequencies) were extracted from each utterance. There were an equal number of feature vectors in the training and test sets.

A decision function was generated for all features in the training set. Using an iterative pattern analysis program, the complexity of the polynomial decision function was systematically increased until all the vectors in the training set were separated into the properly identified languages (100% classification accuracy). When this "tuned" decision function was applied to the evaluation test set, the overall classification accuracy was 84%. The individual language classification scores ranged from 76.8% (American English) to 93.4% (Korean).

Critique. Unlike the studies reviewed thus far, this study specified the languages examined. The individual language scores and the overall accuracy demonstrate the feasibility

of language identification based on acoustic features alone.

It is not clear whether all of the 100 features contributed to the classification performance. Issues such as feature selection, and removal of redundant features need to be examined. Also, the tuning of the decision function on the training set might have attuned the decision function to the idiosyncrasies of the training set, reducing its ability to generalize to a different set of data. A better approach would have been to tune the decision function on a cross-validation or development test set, and then evaluate the tuned decision function on an independent final test set. Also, the relatively small number of speakers per language (five) makes it likely that the system is not truly speaker-independent.

2.1.5 Ives

Using an extended database for the same languages as the previous study, Ives [Ive86] developed an expert system for real-time automatic language identification. The goal of this effort was to develop a set of rules which would minimize the time required for classification.

The extended database consisted of a total of 50 hours of speech from 122 male speakers from each of the following eight languages: American English, Czech, Farsi, German, Korean, Mandarin, Russian and Vietnamese. The speech was originally recorded on magnetic tape, digitized at 10 kHz and low-pass filtered at 5 kHz. Exactly 720 five-second patterns were randomly chosen from each of the 8 languages for analysis, for a total of 5760 patterns. The training and test set subdivisions were not specified.

The classification logic was based on 50 distinguishing features selected by domain experts. An empirical threshold algorithm converted these subjective distinguishing features into objective numerical boundaries or thresholds using the patterns in the training set. These thresholds were used to design a minimum set of nine production rules (all of them were based on the variance of F_0, one on the absolute value of F_0, one on the variance of F_2, one on vowel counts, one on mean "talk windows", and four on power spectral density profiles). Application of this rule set to the test data resulted

in classification scores ranging from 84% (Russian) to 99% (Vietnamese). The overall accuracy was 92%.

Critique. The accuracy figures are impressive, assuming that they were obtained on five seconds of speech per speaker. The amount of speech provided by each speaker is not known, nor are training and test sets used in this study specified. It is also not clear if there was any overlap between the training and test sets. Given that all of the nine production rules were based on F_0, it would be interesting to see the performance of this system on female speakers, who tend to have higher values of F_0 than males. The database used had only male speakers.

2.1.6 Foil

Foil [Foi86] was perhaps the first researcher to report on speech recorded from radio under noisy conditions (the typical signal-to-noise ratio was 5 dB). He imposed an additional constraint that language recognition be made using less than 10 seconds of speech.

The data consisted of 10 hours of speech from each of three unspecified languages, each from a different major language group. (One of them was Slavic, and another was tonal south-east Asian; the third group was not revealed). The training set consisted of 6 hours of speech, the development set of 1.5 hours, and the final evaluation set of 2.5 hours of speech. The number of speakers was not specified.

Two techniques were explored, one based on pitch contours and the other on formant frequencies. The first technique was based on the premise that prosodic features, such as rhythm and intonation patterns which vary from language to language, could be the basis of a powerful language identification technique. In one configuration, a classical quadratic classifier was applied to seven prosodic features extracted from pitch and energy contours in the speech signal. The recognition accuracy on the final test set, using an average of 5 seconds of speech for the identification decision, was 39%. This is only slightly better than chance, given the 3-way choice among the languages.

A second technique was designed to exploit the frequency of occurrence of characteristic sounds of a language by using formant frequency values and locations to represent the sounds. In this configuration, a k-means clustering algorithm determined the 10 best formant vector clusters for each language, and a vector-quantization distortion measure was used as the basis for language decisions. The recognition accuracy on the final test set, using an average signal duration of only 4.5 seconds, was 64%, with a rejection rate of 11%.

Critique. Any evaluation of the results of this study should take into account the noisy data used. The inclusion of a development test set, that was used to provide feedback for the algorithm development process, seems to have helped in "fine-tuning" the features used. The higher score (64%) achieved using formant clusters suggests that formant frequencies are better features than pitch contours when dealing with very noisy speech. Also, given that the languages were from different language groups, it is likely that there are wide phonological differences between them, which might have helped in the identification process.

2.1.7 Goodman et al.

Goodman et al. [GMW89] enhanced Foil's formant extraction technique for language identification by modifying and adding parameters, improving the classifier and reducing its channel sensitivity. A new formant peak-picking algorithm was devised that performed well even with very noisy speech. The original formant vector was augmented with log amplitude values at the formant frequencies, and time-difference terms measuring the formant transitions between significant phonetic events in the language. An improved voiced/unvoiced decision algorithm significantly reduced the number of false voicing errors. A k-means clustering algorithm similar to the one used by Foil was used to determine the 60 best formant-vector clusters for each language. The decision strategy was improved by the use of a weighted Euclidean distance measure instead of a Euclidean distance measure.

The data consisted of a large (9.6 hours), noisy (signal-to-noise ratio: 9 dB), database of six languages, with 2.92 hours of speech in the training set, 2.78 hours in the development set and 3.9 hours in the final test set. The final evaluation was done on a larger database of four different language sets, including this six-language set, the original three-language set used by Foil, and two other geographical subsets. The identity of the languages used was not revealed.

The recognition results were superior to the earlier algorithm in all four language sets (percentage values were not specified). The error rate on Foil's original three-language set was reduced by more than 50%. A significant result was the insensitivity of the recognition accuracy to the signal-to-noise ratio, indicating the robustness of the formant peak-picking algorithm.

Critique. This study achieved significant improvements over Foil's results, and also used several different sets of languages. The robustness of the formant peak-picking algorithm in noise is an important result, since most communication channels are characterized by low signal-to-noise ratios. The lack of information regarding the languages used and the actual classification results precludes comparisons of these results with other work.

2.1.8 Sugiyama

Sugiyama [Sug91b, Sug91a] proposed two language identification algorithms that were based on vector quantization and used acoustic features of the speech signal such as LPC coefficients, autocorrelation coefficients and delta-cepstral coefficients.

The data was taken from a multilingual speech database distributed by NTT, Japan [IIK90]. It consisted of 16 sentences uttered twice by 4 male and 4 female speakers in each of 20 languages (American English, Arabic, Mandarin Chinese, Danish, Dutch, English, Finnish, French, German, Greek, Hindi, Hungarian, Italian, Japanese, Norwegian,

Polish, Portuguese, Russian, Spanish and Swedish)[1]. The duration of each sentence was about 8 seconds. Both the training and test sets had approximately the same amount of data: both in terms of number of speakers and duration (around 21 minutes).

The first algorithm was based on standard vector quantization (VQ). Each language, k, was characterized by its own VQ codebook, V_k, generated using the training sentences. In the recognition stage, input speech was quantized by V_k and accumulated quantization distortion, d_k, was computed. The language with the minimum accumulated distortion was the recognized language. Several spectral distortion measures were tried. The best recognition accuracy, 65% using 64 seconds of unknown speech, was obtained using the weighted likelihood ratios.

In the second technique, a universal codebook $U = \{u_j\}$, was generated using all training data. Each language k was characterized by its occurrence probability histogram h_k. During recognition, each input sentence was quantized by U and its occurrence probability histogram, $h(u_j)$, was computed. The language which had the minimum Euclidean distance between h_k and h was the recognized language. As with the standard VQ, several spectral distortion measures were used for training the universal codebook. The best overall recognition accuracy, 80% using 64 seconds of unknown speech, was obtained using the LPC cepstrum distance.

Human listening experiments were also conducted to determine human language classification performance. Five listeners, four Japanese female and one French male, were presented with pairs of sentences and were asked to judge if they were from the same language or not. If they judged it as the same language, they gave it a value 0. If not, a value 1. If the listener knew one of the two languages in the pair, the value was 0. Since all listeners were familiar with Japanese, it was excluded from the data set. A total of 380 ($= 19 \times 19 + 19$) sentence pairs were presented in random order, the 19 same-language pairs being presented twice.

The values assigned to each language pair were normalized by the total number of

[1]Some of the languages had less than 8 speakers, e.g., Hindi and Greek, while others (English and Norwegian) had more than 8.

sentence pairs presented and a confusion matrix of the languages was generated. A cluster analysis based on the confusion matrix revealed that American English and English were very similar, as were Finnish and Swedish. Chinese and Arabic were found to be isolated from the other languages.

Critique. The results of this study are impressive, considering the large number of languages used. This is clearly one of the best reported results in the literature, and it was achieved on a publicly available corpus of speech. The only caveat is that the amount of speech and the number of speakers per language were small. It would be interesting to see the performance of the VQ techniques on a corpus with a larger number of speakers per language.

The individual language accuracies were not specified, so an analysis of the inter-language confusions is not possible. As for the human perceptual experiments, overall identification accuracy or an analysis of listener performance on individual languages would have been informative. The duration of the speech and the listener's linguistic background clearly influence language identification performance. These effects need to be examined in greater depth. Increasing the number of listeners would also increase the reliability of the results.

2.1.9 Savic et al.

Savic et al. [SAG91] reported preliminary work on language identification using HMMs and pitch contours. The data consisted of 10 minutes of read speech in 4 languages: English, Hindi, Mandarin Chinese and Spanish, recorded in a noise-free room. It was digitized at a sampling frequency of 10 kHz and passed thru a 4.5 kHz low-pass filter. The number of speakers per language was not specified.

The system used features associated with two methodologies: hidden Markov modeling (HMMs) and language-specific pitch contours. A voting classifier was used to combine the results from the HMM and pitch contour modules to arrive at a language identification decision.

Each language was modeled by a five-state linear predictive HMM, and it was found that these states roughly corresponded to different articulatory states of the vocal tract. Preliminary results indicated that transition probabilities for the different languages showed considerable inter-language variations, while the transition probabilities within a language were similar. At the same time, some languages had similar transition probabilities for a few states. This implied that those languages shared a similar phonetic structure for classes of phonemes represented by those states.

The pitch contour analysis module looked at tone and intonation changes across languages. The results on this module were very preliminary, consisting of visual inspection of pitch contours to spot differences between languages.

Critique. This study does not provide any quantitative results and appears to be a preliminary publication of research in progress. The voting classifier mentioned in the overview is not described at all and it is not clear how the pitch contour analysis will be integrated with the HMM results to arrive at the overall language identification score. The one-HMM-per-language concept is an interesting one and merits further investigation.

2.1.10 Nakagawa et al.

Nakagawa et al. [NUS92] examined the application of different HMM-based methods to language identification using acoustic features. They compared the performance of 4 methods: VQ (vector quantization), discrete HMM, continuous density HMM, and mixtured Gaussian distribution model.

The data consisted of 750 utterances from each of four languages, English, Japanese, Mandarin Chinese and Indonesian. Fifteen native male speakers of each language produced 50 sentences each. The type of speech was not specified. The average duration of the utterances was 3 seconds. About 300 utterances from 10 speakers in each language were used to estimate the model parameters. About 100 utterances from the remaining 5 speakers in each language were used to test the models.

The speech was digitized at 12 kHz and a 14-th order LPC analysis was performed with a window of 21.3 ms every 10 ms. The 14 cepstrum coefficients were transformed to 10 mel-cepstrum coefficients.

The standard VQ method was identical to that employed by Sugiyama (see Section 2.1.8). The discrete and continuous HMM models were ergodic (full-structured) rather than left-to-right. An ergodic HMM was constructed for each language and the parameters were estimated using the Baum-Welch algorithm. For the incoming speech, the likelihood of each language was accumulated frame-by-frame. At the end of the utterance, the language with the maximum accumulated likelihood was taken as the system response. HMMs with 2, 3, 5, 7 and 10 states were evaluated. The number of mixtures was set to 1 for all HMMs. For the Gaussian distribution model, the number of mixtures was set at 1, 2, 4, 8 or 16.

The results using the continuous HMMs and mixtured Gaussian models (both 81.1%) were superior to those obtained by the VQ (77.4%) and discrete HMMs (47.6%). The addition of two separate enhancements to the continuous HMMs—duration modeling and dynamic mel-cepstrum features—improved the identification performance to 83.5% and 86.3% respectively.

The addition of 5 more languages, French, German, Korean, Malay and Russian to the data set resulted in a drop in the identification performance of the continuous HMMs from 81.1% to 48%. Pairwise classification between English and Japanese using the continuous HMM with 5 states resulted in an identification accuracy of "more than" 95%.

Critique. The results of the 4-language study are impressive considering the short duration of the utterances (3 seconds). This study has been comprehensive in its examination of different statistical models for language identification.

The sharp drop in performance with the addition of more languages indicates that further experimentation with the model parameters (i.e. increasing number of mixtures and states) may be required. The absence of female speakers in the data set also limits

the generality of the results.

2.1.11 Kwasny et al.

Kwasny et al. [KKWE92] reported preliminary results on a two-speaker two-language identification task using acoustic (raw speech waveform) features and neural network classifiers. The data consisted of 12.5 second samples of read speech in English and French from two bilingual speakers. Two different samples were recorded for each language for each speaker. The speech was recorded in an anechoic chamber resulting in 16-bit samples at 24 kHz. Five band-pass filters were used to separate the signal into bands which were low-pass filtered and decimated by a factor of 200.

The 12.5 second samples of the two speakers were divided into training and test samples. The recognition task was multi-speaker, i.e., different utterances from the same speaker were used in the training and test sets. Each training sample was processed into 371 overlapping 750 ms segments of speech each of which produced 360 numeric values of frequency information across four frequency bands (90 samples in each band). This information was fed to a neural network classifier that performed at 73.7% correct on the test samples. The trained network was then evaluated on varying durations of windows and performance was measured according to a majority vote. This evaluation was done on the combined set of training and test samples. Perfect classification of English was achieved using 2.23 seconds of speech, while French samples needed 3.73 seconds to achieve 100% performance.

Critique. While the use of speech waveform features alone for language identification is novel, this study is preliminary. Clearly, the results do not reflect language identification performance, given the number of speakers in each language (2), the small amount of training and test data, and the fact that the results were achieved on the combined set of training and test samples.

2.2 Summary

There have been only fourteen studies in automatic language identification published in English over the past two decades. The data have spanned the range from phonetic transcriptions of text, laboratory-quality speech, to telephone and radio speech. The number of languages has varied from three to twenty. The approaches to language identification have used "reference sounds" in each language, segment- and syllable-based Markov models, pitch contours, formant vectors, acoustic features, and just raw waveform features. A variety of classification methods have been tried, including HMMs, expert systems, VQ, quadratic classifiers and artificial neural networks.

While the performance figures of some of the studies might look impressive in isolation, meaningful comparisons across studies is not possible, for the following reasons:

- Many of the studies represented classified or sensitive research, so experimental details (e.g., languages used) are often not described.
- There is no common, public-domain database (cf. TIMIT) with which to evaluate different approaches to automatic language identification.

Given the growing importance of automatic language identification services in a rapidly shrinking global community, it is imperative that basic research be conducted in this area. Many of the techniques employed in the studies reviewed in this chapter have produced promising results. They include statistical methods such as HMMs (Nakagawa et. al) and VQ (Sugiyama) as well as a combination of acoustic, phonetic and prosodic features and pattern classifiers. These techniques need to be re-visited, preferably using a large corpus of multi-lingual speech, containing speech produced by several speakers. The availability of such a corpus in the public-domain would enable the proliferation and evaluation of different approaches to the problem. In the next chapter, the development of a ten-language telephone speech corpus, designed specifically for research in automatic language identification, is described in detail. This corpus has since been placed in the public domain.

Chapter 3

Speech Corpus Development

Research in automatic language identification requires a large corpus of multi-lingual speech data to capture the many sources of variability within and across languages. These include variability due to speaker differences (e.g., age, gender, dialect), microphones, telephone handsets, communication lines, background noise and the language being spoken. It is also important that the corpus contain a wide variety of speech from each speaker, ranging from fixed-vocabulary utterances to natural, continuous speech. This makes it useful for both content-dependent and content-independent language identification.

The availability of such a corpus in the public-domain would enable researchers to study languages and to develop, evaluate and compare multi-language recognition algorithms. Unfortunately, there was no such corpus of data available when I began this research. Consequently, I spent a considerable amount of time and effort in collecting and developing multi-lingual speech corpuses. This chapter describes two corpuses that enabled the research reported in this dissertation: (i) a four-language corpus of high-quality (laboratory microphone) speech, and (ii) a ten-language corpus of telephone speech. Both these corpuses contain a mix of natural continuous speech and fixed-vocabulary utterances. These corpuses have since been distributed to several research sites in the United States and abroad, as shown in Table B.1 in Appendix B.

3.1 Four-language High-quality Speech Corpus (OGI_HQ)

3.1.1 Motivation

High-quality speech is recorded in an acoustically controlled environment, and is digitized at a high sampling rate, usually in the 14 kHz to 20 kHz range, which preserves the high-frequency information associated with obstruents (stops, fricatives and affricates) in speech. For my initial research into automatic language identification, this was ideal, since it allowed me to examine the acoustic, phonetic and prosodic differences between languages without being hampered by low signal-to-noise ratios (SNR), channel distortions or unintelligible speech.

The number of languages in this corpus was limited to four. The goal was to examine the issues and problems involved in language identification on a small and manageable body of data, before tackling a larger set of languages.

3.1.2 Language Selection

The choice of the four languages: American English, Japanese, Mandarin Chinese and Tamil, was based more on the availability of native speakers in the Portland metropolitan area than on linguistic considerations. American English and Tamil are stress languages, Japanese is a pitch-accented language, while Mandarin Chinese is a tonal language. The four languages belong to different language families. American English belongs to the Germanic group of Indo-European languages, Japanese is considered by many linguists to be a member of the Altaic family, while others believe it is an Isolate, i.e., belonging to no known language family[Cry87]. Mandarin Chinese is a Sino-Tibetan language, and Tamil is a Dravidian language widely spoken in Southern India and some parts of South-east Asia.

3.1.3 Speakers

Twenty native speakers of each language[1] recorded their voices in our laboratory and were each paid $5 for their participation. The ages of the female speakers ranged from 15 to 70 years while those of the male speakers ranged from 18 to 71 years. All the speakers had spent a major part of their childhood or youth in their native countries. There were approximately equal numbers of male and female speakers in each language. Speaker information for this corpus is listed in Appendix B.2.1.

3.1.4 Recording Protocol

Each speaker produced 20 different speech samples with a few exceptions (see Tables B.2 through B.5 in Appendix B). The speakers were asked to (i) speak 15 conversational sentences on any topic(s) of their choosing, (ii) ask 2 questions of their choice, (iii) recite the days of the week, the months of the year, and the numbers 0 through 10.

The intonation contours for declarative sentences are different from those for questions in many languages. For example, in English, declarative sentences have a falling contour, while questions have a rising contour[GJD87]. The two questions were included in the corpus to capture such intonational differences within languages.

3.1.5 Data Acquisition

The speech was recorded using a Sennheiser HMD 224 noise-canceling microphone, low-pass filtered at 7.6 kHz and sampled at 16 kHz at 16-bit resolution. The recording was done in a carpeted laboratory with a background of computer disk and workstation cooling-fan noise. Each utterance was played back and speakers had the option of re-recording an utterance if it was not considered satisfactory either by them or by the recording supervisor.

[1]There were 24 speakers in Mandarin Chinese and 25 speakers in Japanese.

3.1.6 Corpus Development

Since the data were collected under controlled conditions, there was no need to check for illegal or bad utterances. As such, development of the corpus was restricted to providing broad phonetic transcriptions to a selected subset of the utterances. The speech was automatically segmented into 7 broad phonetic categories using a baseline version of a broad classification and segmentation algorithm (see Chapter 4). The seven broad phonetic categories were: (i) vowels (VOC), (ii) fricatives (FRIC), (iii) stops (STOP), (iv) closures (silence or background noise) (CLOS), (v) pre-vocalic sonorant (PRVS), (vi) inter-vocalic sonorant (INVS), and (vii) post-vocalic sonorant (POVS). The segmenter output was then corrected by trained human transcribers using an interactive display program described in [FPC92]. These broad phonetic transcriptions were used to train and evaluate later versions of the segmentation algorithm.

3.1.7 Corpus Statistics

Not all speakers provided all 20 utterances. Five speakers in Tamil and one in Japanese recorded only 10 utterances each. At the same time, two other speakers in Tamil and one in Japanese recorded a few extra utterances. Table 3.1 displays the salient features of the corpus. The column headings *#M/#F*, *RS/U*, *AvS/U* and *AvUD* represent the number of males and females, range of number of segments per utterance, average number of segments per utterance, and average utterance duration, respectively. The average duration of the utterances in the entire corpus was 5.5 seconds. Figures 3.1 (a) through (g) display the average frequency of occurrence (per second of speech) of the seven broad phonetic categories in the corpus. The low frequencies of occurrence of fricatives in Tamil, of stops in Mandarin, and of post-vocalic sonorants in Japanese suggest that *frequency of occurrence* might be a valuable discriminating feature between the languages. Such features are examined in detail in Chapter 4.

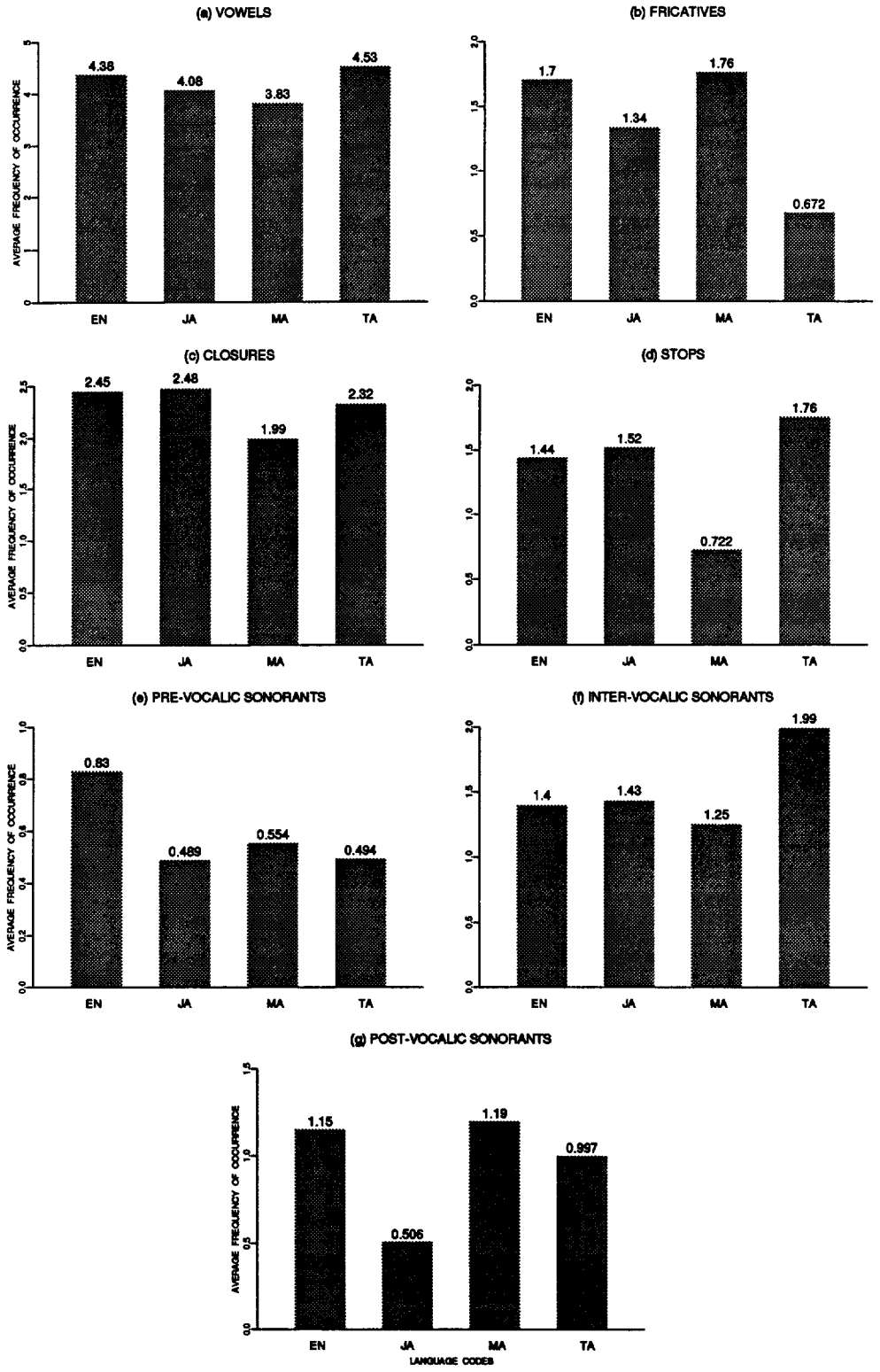

Figure 3.1: Four-language Corpus: Histograms of the 7 Broad Phonetic Categories

Table 3.1: Salient Features of the the Four-language Speech Corpus.

Language	*#Spkrs*	*#M/#F*	*#Utts*	*RS/U*	*AvS/U*	*AvUD*
English	20	10/10	400	15–151	59.1	4.6
Japanese	25	12/13	453	13–153	68.2	6.1
Mandarin	24	14/10	470	11–130	51.1	5.0
Tamil	20	11/09	398	16–155	71.4	5.8

3.1.8 Training and Test Sets

3.1.8.1 Segmentation and Broad Classification

Five utterances from each of 16 speakers per language were used to train and test the segmenter. The training set had 50 utterances from 10 speakers (5 male and 5 female) from each of the 4 languages, for a total of 200 utterances. The development test set had 10 utterances from a different set of 2 speakers (1 male and 1 female) from each language, for a total of 40 utterances. The final test set had 20 utterances from yet another set of 4 speakers (2 male and 2 female) from each language for a total of 80 utterances. The average duration of the utterances in the training set was 4.7 secs and that of the test sets was 5.7 secs. The speakers and utterances used from each language are listed in Appendix B.2.2.

3.1.8.2 Language Classification

Experiments in language identification also used a training set, a development test set, and a final test set. The training set contained 12 speakers from each language, with 10 or 20 utterances per speaker, for a total of 930 utterances. The development test set contained a different group of 2 speakers per language with 20 utterances from each speaker, for a total of 160 utterances. The final test set had 6 speakers per language, with 10 or 20 utterances per speaker, for a total of 440 utterances. The average duration of the utterances in the training set was 5.1 seconds and that of the test sets was 5.5 seconds. The speakers and utterances used from each language are listed in Appendix

B.2.3.

3.2 Ten-language Telephone Speech Corpus

3.2.1 Motivation

There were several problems with the acquisition of a multi-language high-quality speech corpus. The collection process was slow and required considerable human supervision. Addition of new languages to the corpus was heavily dependent on the availability of native speakers of that language in the vicinity of the recording laboratory. Moreover, the usefulness of an automatic language identification system trained only on high-quality speech recorded in a laboratory is rather limited.

In the real world, an automatic language identification system is more likely to be used over some form of communication channel, e.g., a telephone line, which is characterized by low bandwidth, channel distortion, microphone variability and low SNR. If the system is to perform accurately under these conditions, it needs to be trained on speech recorded under these conditions. It was therefore decided to collect speech in many languages over commercial telephone lines. There are several advantages of telephone speech data collection over that of high-quality speech.

- The collection process can be easily automated. Once the recording protocol and equipment are set up, speech data can be collected very rapidly with minimum human supervision.
- Long-distance telephone networks provide access to speakers of different languages spread over a wide geographical area.
- Addition of a new language requires contact with just *two* native speakers (one to pre-record the instructions and prompts, and one to verify them) and a short advertising campaign before data collection in that language can begin.

3.2.2 Language Selection

The ten languages currently in the corpus, English, Farsi (Persian), French, German, Korean, Japanese, Mandarin Chinese, Spanish, Tamil and Vietnamese, were selected based on a combination of linguistic considerations and the availability of native speakers in the United States.

The languages represent a range of unrelated languages (e.g., Vietnamese, Tamil, German) as well as languages from the same sub-family (e.g., Germanic languages such as English and German, Romance languages such as French and Spanish). The languages also include various prosodic features, e.g., Mandarin Chinese and Vietnamese are tonal languages, Japanese uses pitch-accents and syllabic mora. In addition to their linguistic characteristics, the languages represent important geographic and political regions, and many speakers of these languages can be found relatively easily in the U.S.

Since most approaches to language identification rely on discriminators based on patterns of sounds and sound classes, it is important that the corpus include pairs of languages that are phonologically similar and others that are quite distinct. For example, syllable patterns of Vietnamese and Chinese are similar, basically consonant-vowel (CV) or consonant-vowel-consonant (CVC) patterns, with a relatively limited consonant repertoire, and with a characteristic tonal contour associated with each syllable. In contrast, German and English have relatively elaborated syllable structures, potential clusters of half a dozen consonants between vowel nuclei, and no distinctive tonal contrasts at the syllable level. From the point of view of automatic language identification, Chinese and Vietnamese should be more confusable based on phonological sequences, and Chinese and German should be less confusable.

3.2.3 Data Acquisition

3.2.3.1 Collection Campaign

Speaker participation was promoted under a "donate your voice to science" theme. Requests for callers were posted on several university bulletin boards and national computer

network newsgroups. In addition, a press release describing the research project and the need for volunteers resulted in newspaper and radio coverage. A toll-free telephone number, open round-the-clock, was provided.

3.2.3.2 Call Format

A touch-tone phone was needed for the call. Callers received a brief greeting in English followed by a prompt, in each language, to select a language by pressing a digit from 0 through 9. All subsequent instructions and prompts were given in the target language. The objective was to reduce the number of crank calls by non-native speakers. The instructions and prompts in each language were recorded by a native speaker of that language.

3.2.3.3 Recording Equipment

Speech was collected using a Gradient Technology Desklab connected via a SCSI port to a Sun 4/110 workstation. The device was programmed to answer the telephone, play digitized files in each of the ten languages requesting the speech samples, and digitize the callers' response for a designated period of time. Speech was sampled at 8000 samples per second at 14 bit resolution. As the Desklab did not have any automatic gain control mechanism, a fixed recording gain (10) was used.

3.2.3.4 Recording Protocol

The recording protocol was designed to obtain (a) speech samples that create well-defined, useful vocabularies, (b) topic-specific descriptions, and (c) samples of elicited free speech. Small useful vocabularies included

- language names: the speakers' native language and the language they spoke most of the time
- the days of the week

- the numbers 0 through 10

Topic-specific descriptions were obtained by asking callers to describe

- some aspect of their home-town that they liked
- the climate in their home-town
- the room that they were calling from
- their most recent meal

The callers' responses to each of the questions/prompts above were recorded for a fixed length of time (see Appendix A). Elicited free speech was obtained by asking callers to speak for 1 minute on any topic of their choice. Several hints and suggestions about the possible topics were provided: recite a poem, make up a story, describe your favorite sport or hobby, etc. They were then given 10 seconds to organize their thoughts before the actual 1 minute recording. This was done to minimize the number of long pauses and false starts in the free speech. Each speaker contributed 9 utterances, a total of approximately 126 seconds of speech to the database. The duration of each call was approximately 5 minutes. A transcript of the actual prompts and instructions used in the recording protocol is given in Appendix A.

3.2.4 Corpus Development

Development of the corpus was divided into two phases. Phase I, performed by native English speakers, consisted of

- **preliminary verification**: listening to each utterance and deleting prank or invalid calls (hangups)
- **chopping**: removing excess background noise at the beginning and end of each utterance
- **evaluation**: making several judgments about the quality and type of speech

- **broad phonetic transcriptions**: providing time-aligned broad phonetic labels to a subset of the utterances.

Phase II involved

- **verification and evaluation** of the utterances by native speakers of the individual languages
- **orthographic transcriptions** of each utterance
- **time-aligned phonemic transcriptions** (provided by trained human transcribers) of a subset of the utterances for each language

The orthographic and phonemic transcriptions do not have any bearing on the results of this dissertation. As such, these steps will not be discussed any further in this dissertation, apart from mentioning that the Phase II transcriptions are to be used in future work on automatic language identification (see Chapter 6).

The objective of processing the calls in two phases was to obtain a quick preliminary estimate of the number of potentially valid calls in each language, from Phase I. This estimate was helpful in determining when to stop the data collection. The chopping and preliminary evaluation by native speakers of English was fast enough to keep pace with the flow of incoming calls. Data collection was stopped when there were at least 100 potentially valid calls in each language.

3.2.4.1 Phase I

Phase I tasks were carried out by trained laboratory assistants who are native speakers of English. An interactive graphics program was used to display the waveform, play selected portions of the utterance, and to log information into a text file. The speech software tools used in the development of this corpus are described in detail in [FPC92].

Preliminary Verification and Evaluation. Each utterance was processed as follows:

- The utterance was chopped, if necessary, to remove the excess noise and/or silence flanking the speech. Care was taken to include at least 300 ms of "silence" before and after the speech. Audible lip-smacks and breath noise were always retained.
- Judgments were made about the quality and content of speech in each utterance. The listener noted the occurrence of any of the following:
    1. American or British accents (applicable to English calls only)
    2. excessive breath noise
    3. speech cut off at the beginning
    4. speech cut off at the end
    5. environmental noise
    6. caller did not follow instructions
    7. caller not a native speaker
    8. read speech
    9. spontaneous speech
    10. extraneous speech
    11. speech in non-native language
- A set of automatic measurements was made on the utterance. These include its duration, the minimum and maximum sample values, the dc offset, and 10th and 90th percentile of the power (in dB) measured over 10 ms windows in the utterance.

Considering that the verification was done by native English speakers, a few of the above judgments for languages other than English were not very accurate. For example, a native English speaker who does not know Korean has no way of accurately detecting that a Korean caller "did not follow instructions" for the "home-town climate" prompt unless that caller's response was obviously incorrect (e.g., laughing, singing, making animal noises). Similarly, the "caller not a native speaker" comment for languages other than

English was made only if the speaker admitted to being a non-native speaker in response to the "native language" question. A more accurate determination of the number of non-native speakers in each language was made during Phase II, by native speakers of each language. "Extraneous speech" refers to background speech produced by someone other than the caller.

The laboratory assistants were trained to recognize the fixed vocabularies in each language and were able to detect incomplete responses and non-standard pronunciations of the days-of-the-week and the numbers.

In addition to these utterance-specific comments and measurements, the following "global" judgments were made after listening to all utterances of a call:

1. gender (male, female and unknown);
2. age (child, adult);
3. connection quality (poor, average, good); and
4. speaker intelligibility (poor, typical).

Broad Phonetic Transcriptions. As in the case of the four-language high-quality speech corpus, time-aligned broad phonetic transcriptions were provided to selected utterances in each language. The transcription process was semi-automatic, i.e., the output of a segmentation algorithm (described in Chapter 5) was corrected by trained transcribers with the aid of an interactive display program.

3.2.4.2 Phase II

Verification of the utterances by native speakers of the individual languages included (i) confirming that each utterance was in fact spoken by a native speaker of that language, (ii) verifying that the caller followed the instructions for that utterance, (iii) judging the callers' accents and dialects, and (iv) providing a brief description of the topic of each caller's one-minute elicited free speech response.

3.2.5 Current Status

We received a total of 2490 calls over a period of eight months. Of these, 1044 calls were in English, with an average of 160 calls in the remaining 9 languages. On the average, 22.0% of the calls were rejected in each language, mainly because of hangups. A total of 1987 calls (about 43 hours of speech), 868 in English, and an average of 122 calls in the remaining 9 languages, were judged as useful after verification by native speakers. Table 3.2 displays the distribution of "raw" calls, the number of usable calls that resulted from those raw calls, and the average amount of speech (in seconds) per caller, for all the 10 languages. Note that each call can produce a maximum of 126 seconds of speech. The lower numbers in column 6 of Table 3.2 represent the amount of speech actually obtained before the caller decided to terminate the call.

Broad phonetic transcriptions have been provided for 2 utterances per call for the first 25 valid calls in each language (total of 500 utterances). These transcriptions were used in the development of a broad phonetic category segmentation algorithm (Chapter 5).

Table 3.2: Distribution of Calls across 10 languages

Language	*Raw Calls*	*Usable Calls*	*#Utts*	*Avg #Utts/Call*	*Avg Secs/Call*
English	1044	868	7991	9.2	89.7
Farsi	154	115	993	8.6	79.6
French	149	122	1082	8.9	85.4
German	157	118	1059	9.0	87.7
Japanese	147	107	930	8.7	79.5
Korean	149	112	905	8.1	71.1
Mandarin	186	141	1103	7.8	66.4
Spanish	150	128	1150	9.0	86.0
Tamil	194	149	1189	8.0	67.8
Vietnamese	159	127	1023	8.1	69.8

3.2.6 Corpus Statistics

3.2.6.1 Speaker Statistics

The distribution of calls by age and gender judgments is provided in Table 3.3. The "Unk. Gen." column contains the number of speakers whose gender could not be accurately determined by the verifiers. The ratio of male to female speakers was roughly 2.3:1 over all the 10 languages, and ranged from 1.7:1 for German to 6:1 for Tamil.

Table 3.3: Distribution of Calls by Gender and Age Judgments

Language	*Calls*	*Males*	*Females*	Unk. Gen.	*Adults*	*Children*
English	868	595	269	4	853	15
Farsi	115	92	23	0	114	0
French	122	85	37	0	122	0
German	118	74	44	0	118	0
Japanese	107	69	38	0	107	0
Korean	112	84	28	0	111	1
Mandarin	141	89	42	0	141	0
Spanish	128	83	44	1	125	3
Tamil	149	127	21	1	149	0
Vietnamese	127	81	46	0	127	0
TOTALS	1987	1379	602	6	1968	19

3.2.6.2 Utterance Statistics

- 22.9% of the elicited free speech utterances in English were judged to contain read speech
- Figures 3.2 and 3.3 show the average frequency of occurrence (per second of speech) of the seven broad phonetic categories in the utterances of the first 50 calls in each language.
- The average speech rate (number of broad phonetic category segments ÷ utterance duration) for the first 50 calls in each language ranged from 8.02 segments/second

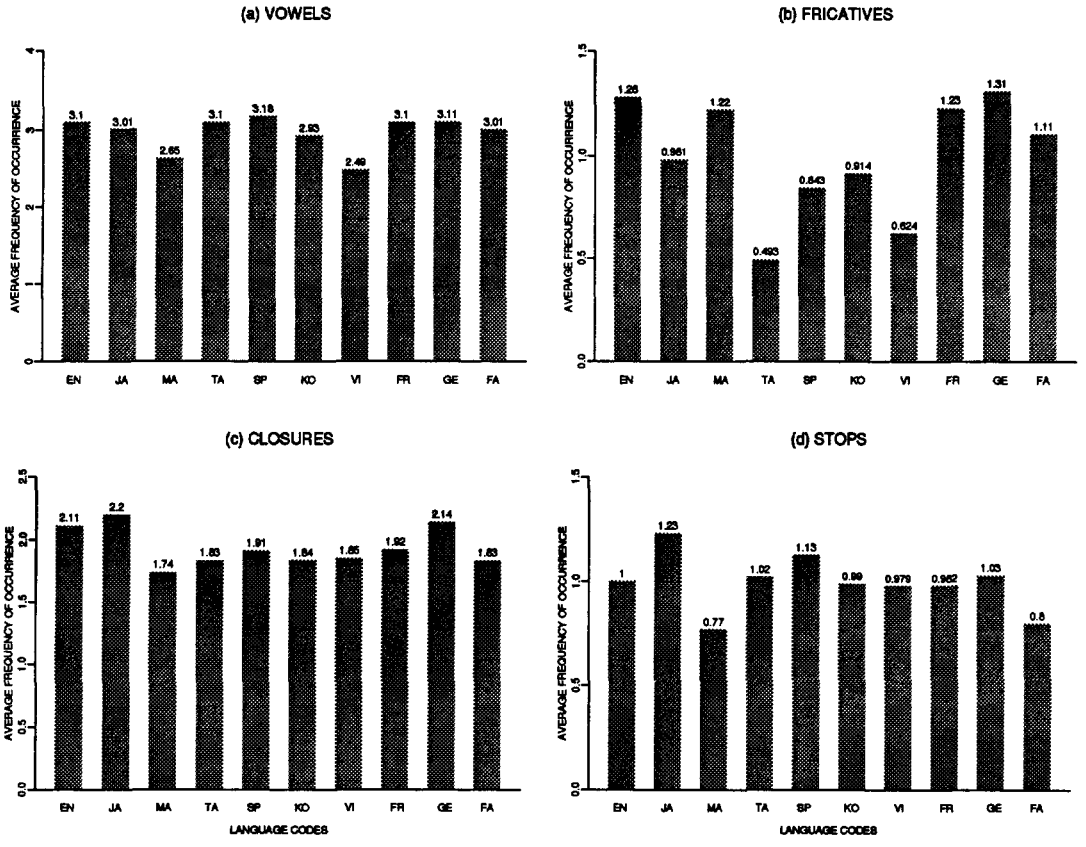

Figure 3.2: Ten-language Corpus: Histograms of VOC, FRIC, STOP and CLOS

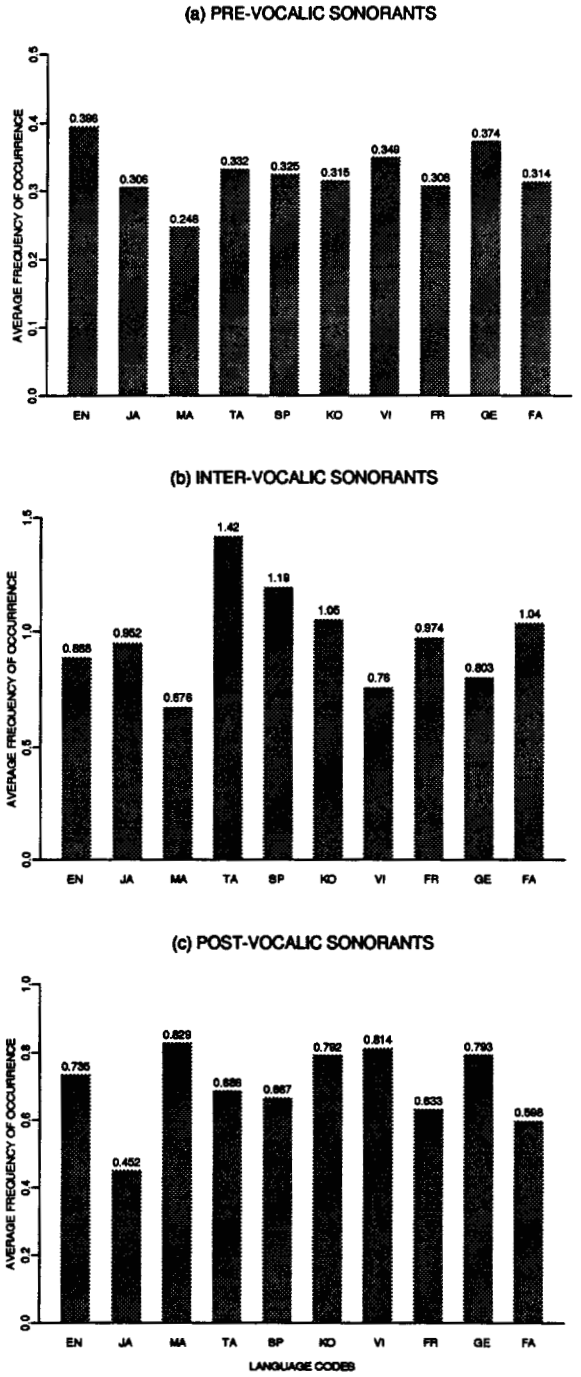

Figure 3.3: Ten-language Corpus: Histograms of PRVS, INVS and POVS

for Vietnamese, to 9.56 segments/second for English, with a median of 8.98 segments/second. These figures include the the long pauses that characterized some callers' 1-minute "story" utterances.

3.2.7 Training and Test Sets

3.2.7.1 Segmentation and Broad Classification

The segmentation algorithm was trained and tested on utterances from the first 25 valid calls in each language. In order to speed up the broad phonetic transcription process, only the following short utterances from each call were used in the segmenter data sets:

- native language
- common language
- home-town liking
- home-town climate
- room description
- meal description

The training set consisted of 300 utterances; 2 per call for 15 calls from each language. The development test set consisted of 100 utterances; 2 per call from a different set of 5 calls in each language. The final test set also consisted of 100 utterances; 2 per call from yet another set of 5 calls in each language. The average duration of the utterances was 4.0 seconds. The actual calls and utterances used are specified in Appendix B.3.2.

3.2.7.2 Language Classification

The language classifiers were trained and evaluated on only the spontaneous speech utterances from the first 90 valid calls in each language. The segmenter data sets described above were completely subsumed by the classifier training set. The training set consisted of 2714 utterances (from 342 males and 158 females); 2-6 utterances per call for 50 calls

in each language. The development test set consisted of 1120 utterances (from 151 males and 49 females); 2-6 utterances per call for 20 calls in each language. The final test set consisted of 1077 utterances (from 139 males and 61 females); 2-6 utterances per call for 20 calls in each language. The utterances ranged in duration from 1 second to 49 seconds with an average of 13.4 seconds. The actual calls and utterances used are specified in Appendix B.3.3.

3.2.8 Human Listening Experiments

How well can human listeners discriminate among these languages from excerpts of speech? To determine human listening performance on excerpts of speech from the 10 languages, 7 female and 4 male monolingual native English speakers were presented with 1-, 2-, 4- and 6-second excerpts of spontaneous speech excised from the 10 languages.

3.2.8.1 Speech Excerpts

The speech excerpts were taken from the 1-minute "story" utterance from each speaker in the training, development and final test sets. Only those stories were chosen that were long enough to provide one excerpt each of four durations: 1, 2, 4 and 6 seconds. Care was taken to ensure that silence constituted less than half of each excerpt. The number of stories chosen from each language was determined by the language containing the least number of stories long enough to provide all four excerpts. Since Korean had only 76 stores satisfying this criterion, 76 different stories from each language were chosen.

3.2.8.2 Experimental Procedure

The experiment was designed such that each subject listened to exactly one excerpt from each speaker, and an equal number of excerpts at each duration from each language. The excerpts were chosen at random, keeping the above constraints in mind. This procedure had the following advantages:

- since each speaker provided only one excerpt, identification of the excerpt was not influenced by familiarity with a particular speaker's voice
- since all excerpts were chosen at random, there was no danger of choosing the longer excerpts from inherently "easy" speakers

The experiment was conducted using an interactive graphics program that played excerpts of speech chosen at random from each of the 10 languages, and maintained a log of subject responses. Following a brief training session, subjects were presented with 760 different excerpts, 19 at each duration from each language. The subjects could listen to each excerpt as many times as they desired. After responding, they were given feedback on every trial. The subjects could also listen to an excerpt *after* making the choice—a feature that was included to aid in the learning process. Each block of 100 trials was considered a session, and the program automatically quit after every 100 trials, to ensure that the subjects did not get fatigued.

3.2.8.3 Results

The average listener performance for each language is shown for the four durations in Figure 3.4. As duration increased from 1 to 2 to 4 to 6 seconds, the average performance over all languages rose from 37.0% to 43.0% to 51.2% to 54.6% respectively. As expected, the subjects identified English with high accuracy at all four durations. Excluding English, the average performance at each duration was 20.7%, 37.4%, 45.8% and 49.7%. Note the relatively high performance on French, German and Spanish—languages that the listeners were most often exposed to, either through courses or by contact with foreign friends. Performance on Farsi, Korean, Tamil and Vietnamese—languages that the listeners were rarely (if ever) exposed to, was very poor.

Analysis of performance by each block of 190 trials revealed little evidence of learning during the experiment. For example, for 6-second excerpts, the average performance on Korean for the first and last 190 trials was 13.5% and 16.7% respectively.

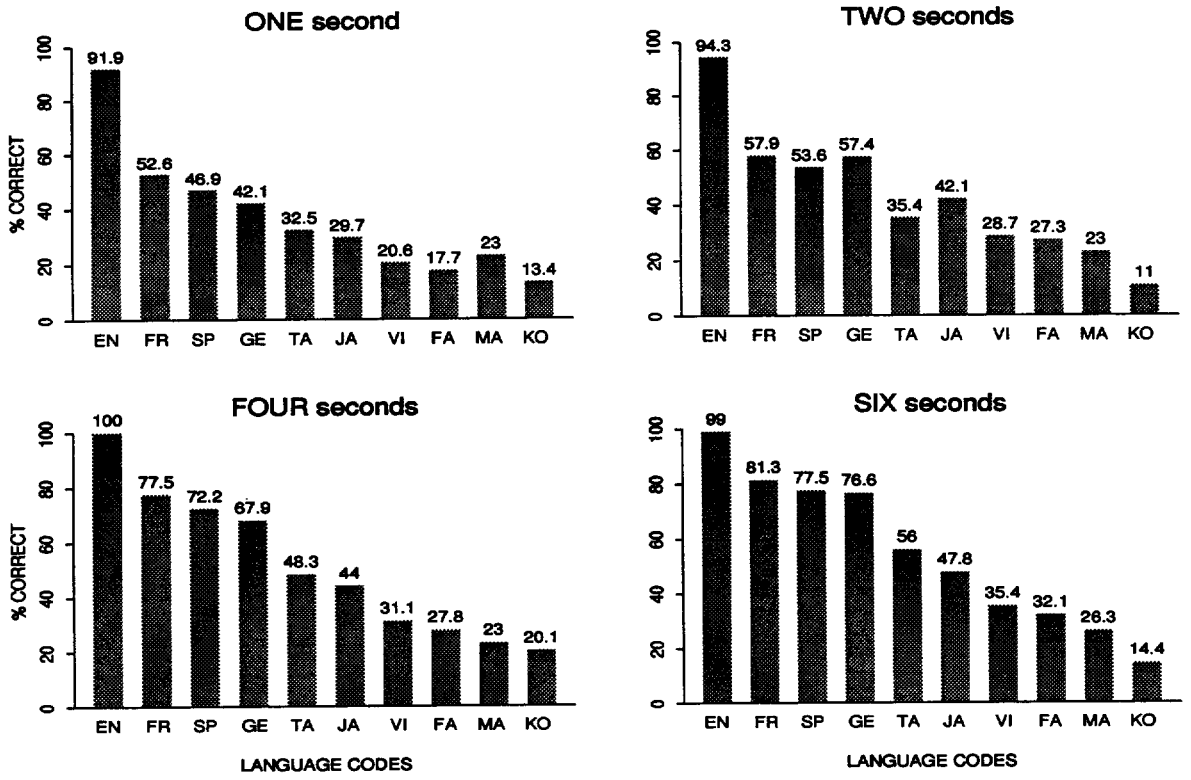

Figure 3.4: Average Listener Performance for the Four Durations

3.2.8.4 Discussion

Since the listeners were trained on speech from all three data sets—training, development test and final test—statistical (or strict) comparisons cannot be made between human and machine identification results (described in Chapter 5). Other differences in the data also exist. The listeners were trained on excerpts ranging from 1 to 6 seconds, while the machine experiments used excerpts ranging from 1 to 49 seconds (Section 3.2.7.2). Further, the listeners were exposed to far fewer utterances than the language classifiers (760 compared to 2714). On the other hand, a human listener has access to other sources of information (such as memory of language sounds overheard in conversations, familiarity with a foreign language, etc.) that are not available to a machine classifier. Therefore, it is not clear that fewer utterances and short durations necessarily handicap human listeners. For this reason, it is interesting to (informally) compare human identification performance with machine results. These listening experiments provide important perceptual benchmarks for automatic language identification. In addition, patterns of confusions between languages provide information about the salient acoustic characteristics that can be useful for automatic language identification.

To determine the effect of additional trials on identification performance, and to examine differences in identification performance for speakers of different languages, a second series of experiments is in progress at the time of writing, using 10 native speakers of English and 2 native speakers of each of the other 9 languages as subjects.

3.3 Summary

In the absence of a public-domain multi-lingual speech corpus, we were left with no choice but to collect and develop our own corpuses to enable our research in automatic language identification. In this chapter, we have described the collection and development of a four-language high-quality speech corpus and a ten-language telephone speech corpus. In addition, we have provided relevant statistics of the corpuses and specified their division into training and test sets for broad phonetic segmentation and language classification.

These corpuses have since been distributed to several research sites within the United States and Europe.

Human listening experiments were conducted on excerpts of spontaneous speech of varying durations excised from the telephone speech corpus. Results indicated that the listeners, who were native speakers of English, performed poorly on unfamiliar languages like Farsi, Korean, Tamil and Vietnamese, regardless of the duration of the excerpts.

In conclusion, it is worth noting that the two corpuses do suffer from one major drawback: they lack samples of free-form conversational speech. It is not immediately obvious that short monologues and extended descriptions are inherently more useful for automatic language identification than truly spontaneous conversational speech. However, it is clear that these corpuses do provide enough data to initiate basic research into automatic language identification. The following two chapters contain a description of our research using these corpuses. The usefulness (or otherwise) of conversational speech for automatic language identification is an interesting research issue that is beyond the scope of this dissertation.

Chapter 4

Automatic Language Identification Using High Quality Speech

This chapter describes the development of a four-language automatic language identification system for high-quality speech. While the development of this system provided valuable insights into the nature of the problem, it is pertinent to stress that this was essentially a feasibility study and therefore did not include detailed parametric analyses of the feature set. The goal of this study was to determine the feasibility of using linguistically motivated features and the classification power of artificial neural networks to perform automatic language identification. Detailed feature analyses were reserved for my research into language identification using ten languages and telephone speech, described in Chapter 5.

An overview of the system is provided followed by detailed descriptions of each of the main stages.

4.1 System Overview

The system uses a neural network-based segmentation algorithm to segment speech into seven broad phonetic categories. Phonetic and prosodic features computed on these categories are then input to a second network that performs the language classification. The following steps transform an input utterance into a decision about what language was spoken.

Data Capture. The procedure for data capture is described in Section 3.1.5.

Signal Representation. A number of waveform and spectral parameters are computed in preparation for further processing. These parameters were empirically derived and were shown to be useful in the segmentation of English letters [Gop90]. The spectral parameters are generated from a 128-point discrete Fourier transform (DFT) computed on a 10 ms Hanning window. All parameters are computed every 3 ms.

The **waveform parameters** consist of estimates of

- *zc8000*: the zero-crossing count of the waveform in a 10 ms window in the frequency range 0–8000 Hz. Zero-crossing is the number of times the waveform crosses the zero line in a 10 ms window. A high zero-crossing count implies high frequency and therefore frication. It is largely independent of the power or amplitude of the signal.

- *ptp700* and *ptp8000*: the peak-to-peak amplitude of the waveform in a 10 ms window in two frequency bands (0–700 Hz and 0–8000 Hz) respectively. Peak-to-peak amplitude is defined as the difference between the highest positive and lowest negative peaks of the original waveform in a 10 ms window. *ptp8000* provides a measure of the waveform envelope and is a very good indicator of silence in clean (high S/N ratio) speech. *ptp700* is computed on the waveform low-pass filtered at 700 Hz. The first formant (the lowest resonant frequency of the vocal tract) is located in the range 0–700 Hz. Since formants are more salient during periodic signals, this parameter gives a fairly good estimate of sonorant intervals.

- *pitch*: the presence or absence of pitch in each 3 ms frame. The pitch estimate is derived from a neural network pitch tracker that locates pitch periods in the filtered (0–700 Hz) waveform [BCVA91].

The **spectral parameters** consist of

- *sda700* and *sda8000*: estimates of averaged spectral difference in two frequency bands (0–700 Hz and 0–8000 Hz). The averaged spectral difference is computed

as the mean squared difference of the spectrum averaged N frames before and N frames after the frame under consideration. N was set to 8 on the basis of experiments performed in [Gop90]. The *sda8000* parameter indicates changes in spectral energy from region to region. Sharp spectral changes indicate the presence of stop bursts. More gradual changes indicate vowel-nasal boundaries. The *sda700* parameter, computed on DFT coefficients below 700 Hz, is used to help determine the onset of voicing, especially vowel onsets.

- *sdf*: fast-change spectral difference in adjacent 9 ms intervals, in the frequency band 0–8000 Hz. This is computed as the mean squared difference of the spectrum in adjacent 9 ms intervals. It has been found useful in detecting abrupt spectral changes, characteristic of stop bursts.

- *cm1000*: the center-of-mass of the spectrum between 0–1000 Hz; in the region of the first formant. This parameter provides a rough estimate of the location of the first formant.

Broad Category Segmentation. Segmentation is performed by a fully-connected, feed-forward, three-layer neural network that assigns 7 broad phonetic category scores to each 3 ms time frame of the utterance. The broad phonetic categories are: vowel (VOC), fricative (FRIC), stop (STOP), closure or silence or background noise (CLOS). pre-vocalic sonorant (PRVS), inter-vocalic sonorant (PRVS) and post-vocalic sonorant (POVS). A Viterbi search, which incorporates duration and bigram probabilities, uses these frame-based output activations to find the best scoring sequence of broad phonetic category labels spanning the utterance.

Language Classification. Language classification is performed by a second fully-connected, feed-forward network that uses 80 phonetic and prosodic features derived from the time-aligned broad category sequence. These features, described in Section 4.3.2.1, are designed to capture the phonetic and prosodic differences among the four languages.

4.2 Neural Network Segmentation

The segmentation and broad classification algorithm described below is a variant of the one developed by Gopalakrishnan for isolated English letters [Gop90]. The training and test sets for segmentation are described in Section 3.1.8.1.

4.2.1 Training the Algorithm

4.2.1.1 Hand-labeling

Both the training and test utterances were hand-labeled with the seven broad phonetic category labels and checked by a second labeler for correctness and consistency. The hand-labeling was semi-automatic. A baseline version of the segmenter was developed using a few utterances hand-labeled completely manually. This segmenter was run on the remaining utterances to be hand-labeled, and insertions, deletions, substitutions and the boundaries of the segments were corrected by human labelers with the aid of an interactive display program.

The three broad phonetic categories PRVS, INVS and POVS were introduced to account for nasals (the phonemes /m/, /n/, and /ng/), semivowels (/l/, /r/, /w/, and /y/) and flaps (/dx/ and /nx/) depending on their position of occurrence with respect to the vowels. For example, a nasal occurring between two vowels was transcribed as INVS, while the /l/ in the word "sleep" was transcribed as PRVS (occurs before the vowel /iy/), and the /n/ in "tent" was transcribed as POVS (occurs after the vowel /eh/ and before the closure /tcl/).

4.2.1.2 Coarse Sampling of Frames

It was not computationally feasible to train on every 3 ms frame in each utterance. Therefore, only a small number of frames were chosen at random from each segment. To ensure that each category was represented by approximately the same number of frames, fewer frames were sampled from the more frequent categories such as vowels and closures. Table 4.1 shows the number of tokens and the number of frames of each category used

to train and test the networks.

Table 4.1: Distribution of Tokens and Frames in the Training and Development Sets

Label	*Training Set*		*Development Set*	
	Tokens	*Frames*	*Tokens*	*Frames*
VOC	943	2829	219	657
FRIC	872	2616	157	471
CLOS	1151	2302	249	498
STOP	981	2942	244	732
PRVS	220	2413	43	465
INVS	848	2544	195	585
POVS	465	2325	114	570
TOTAL	5480	17971	1221	3978

4.2.1.3 Feature Measurements

Feature measurements were derived from the eight parameters described above to provide pattern descriptors sufficient to classify each time frame (3 ms) as one of seven broad phonetic categories. The measurements were taken from a 330 ms window centered on the frame to be classified. Parameter values were sampled from every frame in a 10 frame (30 ms) window centered on the frame, and every 5 frames (15 ms) in two windows spanning 50 frames (150 ms) from each end of the central window. The window sizes were determined empirically. The features were designed to provide detailed information in the immediate vicinity of the frame and less detailed information about the surrounding context. In addition to the 30 samples from each of *zc8000*, *ptp8000*, *ptp700*, *sda8000*, *sda700*, *sdf*, *cm1000* and *pitch*, the spectrum (64 DFT coefficients) *at* the frame to be classified was also included to give a total of 304 features.

4.2.1.4 Network Architecture and Training

The segmentation network was a fully-connected, feed-forward network with 304 input units, 18 hidden units and 7 output units. The number of hidden units was determined

experimentally. Figure 4.1 shows the network configuration and the input features. The inputs to the network were normalized to lie between -1.0 and 1.0. This scaling was required to minimize local-minimum problems [BC89] and to accentuate differences between the features for the seven broad phonetic categories. The normalization procedure for the segmenter features is described in detail in [Gop90].

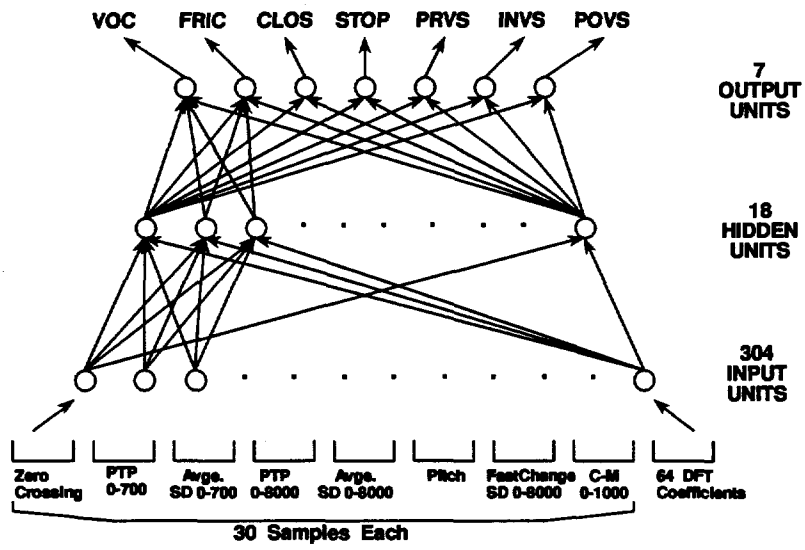

Figure 4.1: Segmentation Network and Input Features

The networks were trained using backpropagation with conjugate gradient optimization [BC89]. Each network was trained on 50 iterations through the training vectors. The trained network was then evaluated on the vectors in the development set to measure generalization. This process was continued and the performance of the network on the test vectors was recorded after every 50 iterations through the training set. The training was stopped when the network had "converged" (attained maximum generalization); convergence was observed as a consistent decrease or leveling off of the classification percentage on the test data over successive sets of 50 iterations. The maximum of the classification percentage on the test data was taken as the final network score. Typically, the networks converged after 500–600 iterations and took 10–20 hours on a Sun 4/75. The above procedure was used in all neural network classification experiments described

in this dissertation.

4.2.2 Segmenter Evaluation

4.2.2.1 Network Performance

Several networks were trained with increasing amounts of training data. All of them were tested on the development test set. Figure 4.2 shows the effect of increasing the number of training speakers per language on the development set classification performance. The highest result (83.6%) was obtained with 10 training speakers per language. Figure 4.2 suggests that performance might further improve with the addition of more training speakers.

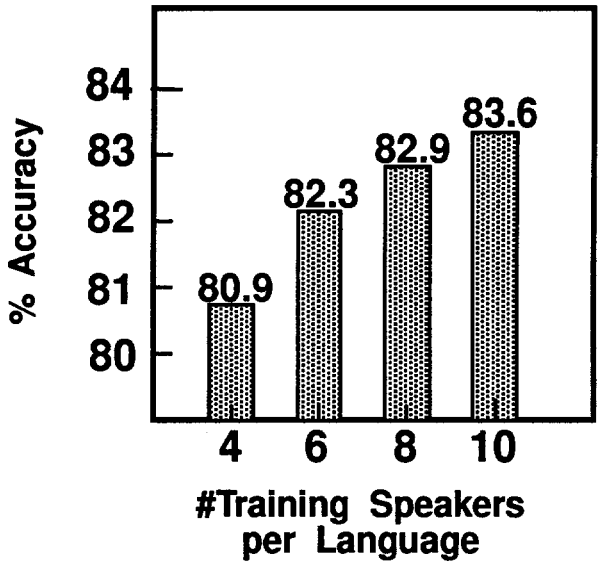

Figure 4.2: Segmentation: Network Performance on the Development Set

4.2.2.2 Performance as an Algorithm

Segmentation performance was evaluated on the 80-utterance final test set using two different scoring procedures. In the first one, the labels output by the segmenter were compared frame-by-frame with the hand-labels, and the percentage of total frames in

agreement was computed. In the second method, a string alignment and scoring program developed by NIST[1] was used. This algorithm treats each segment as a word and measures the number of insertions, deletions and substitutions in the segmenter output with respect to the hand-labels. With the first method, the overall performance accuracy was 85.1%. When scored on the middle 80% and middle 60% of each segment, the accuracy rose to 86.9% and 88.0% respectively, indicating the presence of boundary errors. Table 4.2 shows the results for each broad phonetic category, averaged across languages, for the entire length, and for the middle 80% and 60% of each hand-labeled segment. With the NIST algorithm, the "word" accuracy was 80.4%, with 91.4% correct, 4.0% substitutions, 4.6% deletions and 11.0% insertions.

Table 4.2: Segmentation: Frame-by-frame Scoring on the Final Test Set

Label	*#Frames*	*% Score*		
		Full Segment	*Mid 80%*	*Mid 60%*
VOC	62027	84.4	87.8	90.0
FRIC	17259	91.7	92.8	93.1
CLOS	37138	86.5	86.3	85.1
STOP	5458	76.7	77.4	78.9
PRVS	1963	77.3	78.2	80.6
INVS	12734	85.0	86.4	87.9
POVS	7413	77.5	80.5	83.5
Overall Performance		85.1	86.9	88.0

4.3 Language Identification

Given the broad phonetic sequences, two approaches to language identification were evaluated: (a) sliding window approach, and (b) global features approach. Both these approaches used an artificial neural network for classification, but differed in the method

[1]National Institute of Standards and Technology.

of presenting the inputs to the network. The training and test sets for language classification are described in Section 3.1.8.2.

4.3.1 Sliding Window Approach

In this approach, features from a moving window of N segments were presented to the network at a time. The network produced four language activation scores for each window of N segments. There was an increment of $I(< N)$ segments between successive windows. The four language activation scores were accumulated as the window progressed through the utterance. When the end of the utterance was reached, the language with the maximum accumulated activation score was taken as the system response.

4.3.1.1 Feature Development

For each segment in a window, the following feature measurements were made: (i) broad phonetic label, represented as a vector of length seven, with the appropriate element set to 1.0 and the rest set to -1.0, (ii) the segment duration, and (iii) spectral coefficients averaged over all the frames in S subsegments of the segment. The spectral coefficients were derived from a 7th-order PLP analysis [Her90] on the waveform. This analysis resulted in 8 coefficients (1 energy coefficient and 7 spectral coefficients) for each 3 ms frame of the utterance. Thus, there were a total of $7 + 1 + (8 \times S)$ features from each segment.

4.3.1.2 Network Architecture and Training

The architecture of the language classification network was similar to that used for segmentation. The number of input neurons was determined by the values of N and S. The number of feature vectors (or windows) obtained per utterance was determined by the value of I. The smaller the value of I, the greater the number of windows and hence the greater the number of feature vectors. Different values of N, I, and S were evaluated. It was found that the optimal network performance was obtained for $N = 15$, $I = 1$ (single-segment increment), and $S = 1$ (no subdivisions in each segment). For $N = 15$

and $S = 1$, the network had 240 ($= 15 \times 16$) input neurons, 12 hidden neurons and 4 output neurons, with the number of hidden neurons being determined experimentally. Figure 4.3 displays the network architecture and the schematic of the sliding window of segments. Table 4.3 displays the number of feature vectors in the training and development sets, for $I = 1$. The number of feature vectors in the training set was adjusted so that comparable numbers of vectors were obtained from each language. The relatively low number for Mandarin (6181) was due to the fact that there were fewer segments per utterance on the average in Mandarin than in the other three languages.

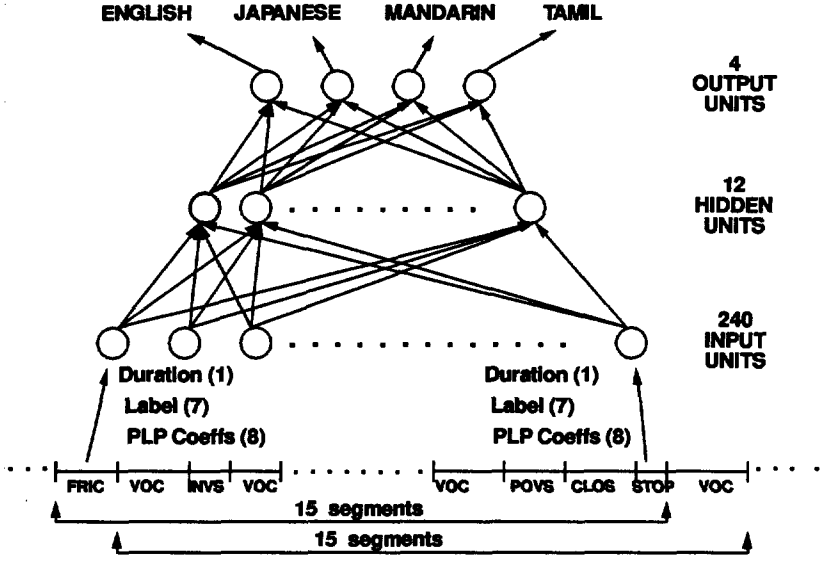

Figure 4.3: Sliding Window Approach: Network Architecture and Input Schematic

As before, the input features were normalized to lie between -1.0 and 1.0. The label vector and the 7 spectral coefficients were already in the required range and did not need any further scaling. The segment duration was normalized as follows: histograms of the durations of the 7 broad phonetic categories across all languages were generated. For each histogram, the upper and lower duration bounds, $maxdur_s$ and $mindur_s$, that contained 90% of the data were determined (s being one of the 7 broad categories). Individual duration values, dur_s, were then normalized to lie between -1.0 and 1.0 using

the following formula:

$$normdur_s = 2.0 \times (\frac{dur_s - mindur_s}{maxdur_s - mindur_s}) - 1.0 \quad (4.1)$$

The energy coefficient of the PLP was normalized in similar fashion using upper and lower bounds generated from a histogram of all PLP energy values within the utterance.

The training algorithm and the procedure for determining maximum generalization were identical to those described in Section 4.2.1.4.

Table 4.3: Sliding Window Approach: Distribution of Feature Vectors for $I = 1$

Language	*Training Set*	*Development Set*
English	8822	1568
Japanese	8822	2226
Mandarin	6181	1391
Tamil	8822	2775
TOTAL	32647	7960

4.3.1.3 Language Identification Performance

The best network performance on the two-speakers-per-language development set was 71.8%. This result represented the overall recognition accuracy of the 15-segment windows. The language identification performance on utterances was determined by accumulating the language activation scores for all windows in an utterance, as explained above. With this method, the language identification performance on the development set was 84.4%, with English 67.5%, Japanese 85.0%, and Mandarin Chinese and Tamil both 92.5%.

However, on the 6 speakers-per-language final test set, the corresponding language identification performance dropped to 61.6%, with English 79.2%, Japanese 56.7%, Mandarin Chinese 39.2% and Tamil 70.0%. Figure 4.4 displays the corresponding confusion matrix and the individual language classification percentages. The row labels represent the actual language and the column labels represent the system response. The one-way

confusions between the languages are striking. Japanese and Mandarin show a tendency to get misclassified as English, while Tamil gets misclassified as Japanese.

label	EN	JA	MA	TA	
EN	95	6	4	15	79.2%
JA	25	51	9	5	56.7%
MA	46	14	47	13	39.2%
TA	2	34	.	84	70.0%

Figure 4.4: Confusion Matrix for the Sliding Window Approach

4.3.2 Global Features Approach

It is possible that the performance of the sliding window approach could have been improved with the addition of more sophisticated features from each segment (e.g., pitch contours). But in keeping with the breadth-first nature of this study, I decided to try an approach wherein the features were computed over the utterance as a whole, rather than in a moving window. Hence the name *global* features.

Since the neural network required all vectors to be of the same length, the global features had to be duration independent, i.e., the feature computation had to yield a fixed number of values regardless of the utterance duration.

4.3.2.1 Feature Development

Several passes were needed through the iterative process of feature development and network training before a satisfactory feature set was obtained. Much of the effort was concentrated on exploratory statistical and linguistic analysis of the languages with the objective of determining the distinguishing characteristics among them. For example, the knowledge that Mandarin Chinese was the only tonal language in the set, led us to design features that attempted to capture the large variation in pitch within and across segments for Mandarin Chinese utterances. Similarly, the presence of sequences of almost

equal-length broad category segments in Japanese utterances led us to design an "inter-segment duration difference" feature. Histograms of the frequency of occurrence of the broad phonetic categories (Figure 3.1) showed that (i) Tamil had the lowest frequency of occurrence of fricatives of the four languages, while Mandarin had the highest, (ii) Mandarin had the lowest frequency of occurrence of stops, and (ii) Japanese had the lowest frequency of occurrence of post-vocalic sonorants.

The final set of 80 features is described below. All the features were computed over the entire length of an utterance and used the time-aligned broad category sequence provided by the segmentation algorithm. The numbers in parentheses refer to the number of values generated.

- Intra-segment pitch variation: Average of the standard deviations of the pitch within all sonorant segments—VOC, PRVS, INVS, POVS (4 values)
- Inter-segment pitch variation: Standard deviation of the average pitch in all sonorant segments (4 values)
- Frequency of occurrence (number of occurrences per second of speech) of triples of segments. The following triples were chosen based on visual inspection of the broad phonetic sequences in each language: VOC-INVS-VOC, CLOS-PRVS-VOC, VOC-POVS-CLOS, STOP-VOC-FRIC, STOP-VOC-CLOS, and FRIC-VOC-CLOS (6 values)
- Frequency of occurrence of each of the seven broad phonetic labels (7 values)
- Frequency of occurrence of all segments (number of segments per second) (1 value)
- Frequency of occurrence of all obstruents (STOPs and FRICs) (1 value)
- Frequency of occurrence of all sonorants (1 value)
- Ratio of number of sonorants to total number of segments (1 value)
- Fraction of the total utterance duration devoted to each of the seven broad phonetic labels (7 values)

- Fraction of the total utterance duration devoted to all sonorants (1 value)
- Frequency of occurrence of voiced obstruents (1 value). An obstruent was considered voiced if more than half the number of frames in it were labeled as voiced by the neural network pitch tracker [BCVA91]
- Ratio of voiced consonants to total number of consonants (1 value)
- Average duration of the seven broad phonetic labels (7 values)
- Standard deviation of the duration of the seven broad phonetic labels (7 values)
- Segment-pair ratios: the ratios $num(AB)/num(AX)$ and $num(AB)/num(YB)$, where A and B are the members of a particular segment pair, X is any segment that can legally follow A, Y is any segment that can legally precede B, and num(S) is a function that returns the number of occurrences of the segment-pair S. The segment-pairs were selected based on histogram plots generated on the training set. Examples of selected pairs: POVS-FRIC, VOC-FRIC, INVS-VOC, etc. (27 values)
- Inter-segment duration difference: Average absolute difference in durations between successive segments (1 value)
- Standard deviation of the inter-segment duration differences (1 value)
- Average distance between the centers of successive vowels (1 value)
- Standard deviation of the distances between centers of successive vowels (1 value)

4.3.2.2 Network Architecture and Training

The network had 80 input neurons, 29 hidden neurons and 4 output neurons. The number of hidden neurons was determined experimentally. All 80 features were normalized to lie between -1.0 and 1.0, using the normalization procedure described in Section 4.3.1.2. The training algorithm and the procedure for determining maximum generalization were identical to those described in Section 4.2.1.4. Table 4.4 displays the distribution of

feature vectors in the training, development and final test sets. Note that each utterance yields one feature vector, unlike the sliding window approach (Table 4.3). As a result, the neural network classification score for each utterance *was* the language identification score for that utterance.

Table 4.4: Global Features Approach: Distribution of Feature Vectors

Language	*Training Set*	*Development Set*	*Final Test Set*
English	239	40	120
Japanese	213	40	100
Mandarin	240	40	120
Tamil	238	40	120
TOTAL	930	160	460

4.3.2.3 Language Identification Performance

Single Utterances. During the feature development phase, the 2 speakers-per-language development test set was used. The classifier performed at an accuracy of 90.0% on this small test set. For final evaluation, the development test set was combined with the original training set to form a 14 speakers-per-language training set. The performance of the classifier on the 6 speakers-per-language final test set was 79.6%. The individual language performances were English 75.8%, Japanese 77.0%, Mandarin Chinese 78.3%, and Tamil 88.0%. This result was obtained with training and test set utterances that were approximately 5.5 seconds long on the average, and contained a mixture of fixed-vocabulary utterances (days-of-the-week, months-of-the-year and digits) and free speech. To make sure that the performance was not being inflated by these fixed segment sequences, the fixed vocabulary utterances in the test set were removed. With just the free-speech test set, the overall performance dropped from 79.6% to 79.5%, an insignificant change.

For the sake of fair comparison between the global features approach and the sliding window approach, a network was trained on global features from the same 12 speakers-per-language training set used in the sliding window approach (Section 4.3.1). This

network performed at 78.0% accuracy on the final test set. This represents a 42.7% reduction in the error-rate compared to the 61.6% accuracy obtained with the sliding window approach. Figure 4.5 displays the confusion matrix and the individual language scores corresponding to the 78.0% result. The one-way confusions seen between Japanese, Mandarin and English, and between Tamil and Japanese with the sliding window features (Figure 4.4) are substantially reduced with the global features. Mandarin displays the largest jump in performance, from 39.2% to 76.7% (a 61.7% reduction in the error-rate). Also, the overall identification accuracy with global features showed no appreciable increase (78.0% to 79.6%) with increase in the training data from 12 speakers to 14 speakers per language.

label	EN	JA	MA	TA	
EN	95	8	10	7	79.2%
JA	16	71	2	11	71.0%
MA	14	5	92	9	76.7%
TA	6	8	1	85	85.0%

Figure 4.5: Confusion Matrix for the Single Utterances

Concatenated Utterances. To observe the effect of training and testing with longer durations of speech per utterance, a series of experiments were conducted in which pairs and triples of utterances from each speaker were concatenated end-to-end (with 350 ms of silence in between the utterances to simulate natural pauses) in both the training and test sets. It is to be noted that the total duration of speech used in training and testing remained unchanged for all these experiments. This concatenation of the utterances was preferred to the time-consuming and labor-intensive task of recruiting more speakers in each language and having them speak longer utterances. Table 4.5 summarizes the performance of the classifier when trained and tested on different durations of speech per utterance. Each row of the table shows the effect of testing on progressively longer utterances for a given training set, while each column of the table shows the effect

of training on progressively longer utterances for a given test set. Not surprisingly, the best performance (89.5%) is obtained when the classifier is trained and tested on three utterances concatenated together. Figure 4.6 displays the corresponding confusion matrix. As expected, performance of all languages increases with duration. The largest reduction in error-rate with respect to the single utterances (Figure 4.5) was obtained for English (80%), while the smallest drop was for Mandarin Chinese (23.0%).

With just the free speech utterances in the test set, the corresponding best language identification score dropped from 89.5% to 88.5%, an increase of 9.5% in the overall error-rate.

Table 4.5: Percentage Accuracy on Varying Durations of Speech per Utterance

Avge. Duration of Training Utts. (sec)	*Avge. Duration of Test Utts. (sec)*		
	5.7	11.8	17.1
5.3	79.6	73.6	71.2
10.6	71.8	86.8	85.0
15.2	67.9	85.5	89.5

label	EN	JA	MA	TA	
EN	40	0	1	1	95.2%
JA	6	28	0	0	82.4%
MA	3	3	35	1	83.3%
TA	0	1	0	34	97.1%

Figure 4.6: Confusion Matrix for the Concatenated Utterances

Error-rates vs. McNemar's Test. It is pertinent at this stage to explain my preference for reporting reduction in error-rates rather than $p-$values from the McNemar's Test, while comparing neural network classification results. In selecting the "optimal"

network configuration (i.e., number of hidden units and the random seed for the initial network weights) through parametric experiments, it is impossible to examine the space of all possible network configurations. There might well be a combination of hidden units and random weight seed that performs better than our chosen "optimal" configuration. Thus, differences between a pair of classification results are caused not only by random factors such as the selection of a particular set of training samples, but also by these complex deterministic choices. For this reason, it does not make mathematical sense to talk about statistical significance ($p-$values) of neural network classification results. Reporting the changes in error-rates provides a more realistic comparison of the results in this case.

4.4 Discussion

Clearly, the features used in the sliding window approach did not generalize well to the larger final test set. The inflated performance on the development set can be attributed to a favorable (albeit random) set of speakers from each language in that set. It should be pointed out that the feature set did not contain any pitch information, which might explain the low recognition accuracy of Mandarin Chinese (39.2%).

The results of the global feature approach suggest that (i) features computed on the utterance as a whole perform better than features in a moving analysis window, and (ii) global features perform better on longer utterances. While it is debatable whether the first observation would still be valid if more sophisticated moving-window features were to be examined, the second observation is certainly valid, given the nature of the feature set. Segment-based statistical features tend to be more reliable with a larger number of segments. Also, it is interesting that a language identification accuracy of 89.5% was possible without using any spectral information in the classifier feature set. All of the features were based on the broad phonetic category segment sequences provided by the segmenter.

Any attempt to use spectral features, derived from utterances of varying length, as

input to a static neural network should address the following questions: how does one account for differences in utterance duration? How does one pick the frames from which the spectral information is sampled? What number of frames is considered "optimal?" If an utterance is shorter than the chosen fixed number, the feature vector will have to be padded with zeros, since the network requires equal-length vectors. Is this acceptable? These questions are addressed in the next chapter.

4.5 Summary

In this chapter, I have described a segmental approach to automatic language identification using four languages and high-quality speech. Neural network-based broad phonetic segmentation of speech was followed by the computation of linguistically motivated features based on the broad phonetic sequences. These features were fed into a second network which classified the languages. It was found that features computed over the entire utterance provide better language identification performance than features from a sliding window of segments.

Despite the fact that the high identification accuracy was obtained on a set of languages with wide phonological differences, the results of this feasibility study are encouraging. With just the information in seven broad phonetic categories, four languages were identified with an accuracy of 89.5% using utterances that were 17.1 seconds long on the average. The next chapter examines the application of this approach to a more difficult task: identification of ten languages using speech recorded over commercial telephone lines.

Chapter 5

Automatic Language Identification Using Telephone Speech

5.1 Introduction

Despite the myriad problems associated with telephone speech—low bandwidth, channel distortion, excessive noise, hand-set variability—it does have one overwhelming advantage: it is the kind of speech we encounter every time we use the phone. Given the widespread use of telephones, a language identification system using telephone speech will be significantly more useful than one trained on clean, laboratory speech.

This chapter describes research in automatic language identification using speech recorded over commercial telephone lines. The speech corpus contained ten languages (Section 3.2) but the basic approach was essentially the same as that for high-quality speech: segmentation of speech into broad phonetic categories, followed by language classification using linguistically motivated features measured on sequences of the broad phonetic categories. The main difference was that spectral and pitch-based features, unexplored in the research using high-quality speech, were evaluated using telephone speech. Further, detailed parametric analyses of the feature set were carried out to determine the features most useful for language identification.

The neural networks described in this chapter were all trained using backpropagation with conjugate gradient optimization [BC89]. In every case, the testing procedure and criterion for stopping training were identical to those described in Section 4.2.1.4.

After a description of the broad category segmentation, the rest of this chapter is devoted to testing the hypothesis that knowledge-based features provide better language identification performance than features with minimal linguistic knowledge encoded in them. The performance of baseline segment-based features is compared to that obtained with features that are either linguistically motivated or derived from exploratory statistical analyses of the segment sequences. Such features include (a) features based on triples and pairs of broad phonetic categories, and (b) pitch variation features. In addition, spectral features, with and without segmentation, are also evaluated. The tasks examined include identification of all ten languages, and pairs and triples of languages with English as the anchor. Where applicable, the effect of reducing the dimension of the feature set through principal component analysis and visual inspection of feature plots is also investigated. The individual feature sets are evaluated on the development test set and a combined set of features is determined. Using this combined set, language identification performance on the final test set is reported.

5.2 Ten-language Telephone Speech Corpus

The ten-language telephone speech corpus consisted of fixed-vocabulary and spontaneous speech in English, Farsi (Persian), French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil and Vietnamese, produced by native speakers of each language. The design and development of this corpus, and the subdivision into training, development and test sets are described in detail in Section 3.2.

5.3 Broad Phonetic Category Segmentation

The broad phonetic segmentation algorithm described here is a variant of the fine phonetic segmentation algorithm developed for telephone speech in English [Rog91]. It is similar to the one described in Section 4.2 for high-quality speech, in the sense that it uses the same seven broad phonetic categories and the same procedure: frame-based neural network classification followed by a Viterbi search using duration and bigram

probabilities. There were, however, two major differences:

- the spectral coefficients were obtained from a PLP analysis of the waveform [Her90] instead of the DFT.
- frames randomly sampled from each hand-labeled segment (Section 4.2.1.2) for training were augmented with frames sampled at the edges of segments. Such edge-sampling has been shown to improve frame classification performance for English phonemes [Rog91].

5.3.1 Why Use PLP?

Conventional linear predictive coding (LPC) analysis [Mak75] models the power spectrum of speech equally well at all frequencies of the analysis band, making it inconsistent with human auditory processing. For example, it has been shown that the human ear is most sensitive in the middle range of the auditory spectrum, and has decreasing frequency resolution above 800 Hz. The PLP analysis of speech [Her90] attempts to mimic the properties of the human auditory system by applying mathematical transformations to the short-term power spectrum produced by conventional LPC analysis, then approximating the new, auditory-like spectrum with an autoregressive all-pole model. The mathematical transformations are, in order: integration over simulated critical-band auditory masking curves, followed by resampling the integrated spectrum in approximately 1 Bark intervals; pre-emphasis by a simulated fixed equal-loudness curve; and compression of the resulting spectrum through the cubic root nonlinearity to simulate the intensity-loudness power law of hearing.

PLP analysis also has the advantage of producing a more compact representation than the DFT. Janssen et al. [JFC91] showed that 8 PLP coefficients per frame worked just as well as 40 DFT coefficients per frame for frame-based phonetic classification of English letters using high-quality speech. Junqua [Jun89] demonstrated the superiority of PLP over standard LPC on an alphadigit recognition task. A comparison of PLP, DFT and cochleagrams [Lyo82, Sla88] on the English letter recognition task by Fanty

and Cole [FC91] showed the superiority of PLP over the other two representations. Creekmore et al. [CFC91] investigated five spectral representations—PLP, DFT and three representations based on LPC—for speaker-independent phonetic recognition using the TIMIT corpus. They found that the PLP outperformed the other representations with less computation and shorter learning times. Chigier and Leung [CL92] found that the PLP produced the lowest error-rates on a 39-phone recognition task using multi-layer perceptrons, for both the TIMIT and NTIMIT [JKBS90] corpuses. In the light of such substantial evidence for its effectiveness, for both high-quality and telephone speech, PLP analysis was the spectral representation of choice for the spectral-based features used in this research.

Unless otherwise specified, the PLP analysis performed in all the experiments in this dissertation was of the 7th-order with a 10 ms window and a 3 ms increment. This resulted in 7 cepstral coefficients and 1 energy coefficient per frame. The cepstral coefficients were weighted with an exponential window so that all coefficients had a similar range for input to a neural network. The energy coefficient was normalized with respect to the minimum and maximum values in each utterance using a formula similar to that in Equation 4.1.

5.3.2 Edge-sampling

Roginski [Rog91] developed a neural network-based fine phonetic classifier for telephone speech for the English alphabet recognition task. He demonstrated the benefits of sampling at the edges of the hand-labeled segments in addition to sampling frames at random within the segments. He identified a set of "critical" phone boundaries that needed to be corrected. He showed that edge-sampling dramatically improved placement of the [b]/[iy] boundary. Overall, the random sampling network placed 46.4% of "critical" phone boundaries within one frame of the hand labels, compared to 58.2% by a network trained on both random and edge-samples. Roginski's experiments and visual inspection of the broad phonetic segmentation output provided compelling evidence of the need to sample at the edges of the hand-labeled broad phonetic segments.

In our experiments, the randomly sampled frames from each segment were augmented with equal numbers of edge-samples from both edges of the segment.

5.3.3 Feature Development

The spectral features consisted of 120 averaged PLP coefficients taken from various points in a 435 ms window centered around the frame to be classified. These features were empirically derived to capture the contextual information in the vicinity of each frame [FCR92]. Figure 5.1 shows the sampling intervals for the PLP coefficients in the context surrounding the frame to be classified. The solid boxes indicate 9 ms and 15 ms intervals over which PLP coefficients are averaged. Dashed boxes indicate 18 ms intervals that are skipped.

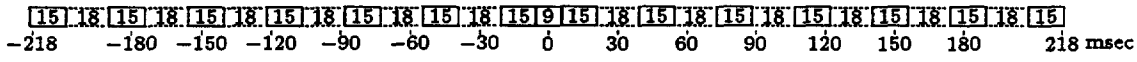

Figure 5.1: Sampling Intervals for the PLP coefficients.

Experiments were conducted to examine the effect of adding waveform-based features such as zero-crossing rate and peak-to-peak amplitude and spectral features such as spectral difference to the 120 PLP-based features. These features produced no improvement in comparison to networks that were trained with just the 120 PLP-based features.

5.3.4 Segmenter Training

As before, the utterances in the training and test sets were hand-labeled with the 7 broad phonetic categories. Table 5.1 displays the distribution of tokens and frames of each broad phonetic category used to train the networks. The numbers in the *Frames* column for both the training and development sets include 3000 edge-sampled frames from each broad phonetic category, the balance consisting of randomly sampled frames. The network had 120 input units, 35 hidden units and 7 output units, with the number of hidden units determined experimentally.

Table 5.1: Distribution of Tokens and Frames in the Training and Development Sets

Label	*Training Set*		*Development Set*	
	Tokens	*Frames*	*Tokens*	*Frames*
VOC	3274	6251	1297	5578
FRIC	1314	6336	473	5353
CLOS	2413	6530	787	5284
STOP	1150	6523	442	5795
PRVS	383	6568	124	4991
INVS	887	6424	390	6037
POVS	648	6462	265	5854
TOTAL	10069	45094	3778	38892

5.3.5 Viterbi Search

The frame-by-frame outputs of the segmenter were converted into a time-aligned sequence of the 7 broad phonetic category labels using a Viterbi search with duration and bigram constraints obtained from the hand-labeled utterances of the training set. In addition, the search used an insertion penalty that penalized all segments, reducing the incidence of short, spurious insertions in the segment sequence. Different values of the insertion penalty were evaluated on the development set using the output scores of the best network. The value that provided the best frame-by-frame comparison score (Section 4.2.2.2) on the development set, 0.035, was chosen.

5.3.6 Segmenter Evaluation

Several networks were trained and evaluated on the development set. The best network performance was 71.6% on the development set. The segmenter was scored on the final test set using the two scoring techniques described in Section 4.2.2.2. With the frame-by-frame comparison scoring method, the performance accuracy was 79.8%. This compares favorably with 85.1% for 4 languages using high-quality speech. When scored on the middle 80% and middle 60% of each segment, the accuracy rose to 82.2% and 83.9% respectively. Table 5.2 shows the performance accuracy for each broad phonetic category,

averaged across languages, for the entire length, and for the middle 80% and 60% of each hand-labeled segment. With the NIST word-alignment algorithm, the performance accuracy was 72.2%, with 81.8% correct, 6.9% substitutions, 11.2% deletions, and 9.7% insertions. In comparison, the word-accuracy with high-quality speech was 80.4%.

Table 5.2: Telephone Segmentation: Frame-by-frame Scoring on the Final Test Set

Label	*#Frames*	*% Score*		
		Full Segment	*Mid 80%*	*Mid 60%*
VOC	80136	76.5	80.4	83.6
FRIC	26562	76.6	79.2	80.7
CLOS	80513	89.4	91.1	91.8
STOP	6853	53.6	54.0	56.0
PRVS	3739	63.5	64.1	64.9
INVS	13922	74.2	75.7	77.4
POVS	10745	69.7	71.8	73.5
Overall Performance		79.8	82.2	83.9

5.4 Language Classification

To the best of my knowledge, all of the literature on linguistic and phonological differences between languages examines differences at the level of phonemes or phones [Cry87, Com90, Mad84]; levels that are beyond the scope of this dissertation. Therefore, visual inspection and statistical analyses of the broad phonetic segment sequences were the main tools used in the selection and development of features for language identification. From this point onwards, the term "segment" will be used to refer to "broad phonetic category segment," unless otherwise specified.

5.4.1 Feature Selection

The segment-based features used in the language classification experiments were motivated by the following objectives:

- to exploit specific known language differences e.g., the tone variations in Mandarin, the "mora" in Japanese, the presence of consonant clusters in English, etc.,
- to determine differences in the general statistics of segment sequences from the ten languages.

Features based on segment sequences can be divided into two categories:

- segment label and duration information, with minimum encoding of linguistic knowledge. The objective was to determine if the neural network could automatically extract the statistical differences among languages given all of the segment information, without any pre-processing.
- features derived from exploratory analyses of segment sequences, such as average duration, ratio of number of occurrences of each broad category, frequency of segment-pairs, segment-triples, etc. The objective here was to identify features that accentuated language differences, with the help of statistical analyses of segment sequences, and to present the processed information to the network.

In the following sections, I test the hypothesis that language classifiers perform significantly better with knowledge-based features than with features that have minimal linguistic knowledge encoded in them. To this end, the performance of a baseline classifier that uses just segment duration and label information is compared with that of classifiers using knowledge-based features derived from pairs and triples of broad phonetic categories. The use of spectral features, both with and without segmentation, is also examined. The experiment using spectral features without segmentation provides yet another baseline result for language identification performance. It illustrates the level of language identification accuracy possible without any segmentation of the speech signal.

To determine if the features being used were redundant, principal component analysis and visual inspection of boxplots (described in Section 5.4.5) were used to reduce the dimension of the feature space.

While the classification performance of pairs of languages and other language combinations smaller than ten might indeed be informative, it was decided to focus on the ten-language task during the feature development phase, since it was the most challenging of all the classification tasks with this corpus. Therefore, where feasible, the experiments examined classification of all ten languages with a single network. In cases where using input features from all ten languages at once was computationally infeasible, or where a mere concept proof was considered more important than detailed performance figures involving all languages, the classification involved just English and Japanese (a single 2-language network) or the nine language pairs with English as the *anchor* language (nine separate 2-language networks). Section 5.4.10.2 examines classification performance of language pairs and other smaller language combinations on the final combined feature set.

5.4.2 Baseline Experiment - Segment Duration and Label

This experiment was designed to determine the language discrimination ability of an artificial neural network when presented solely with information about the duration and category scores of each broad phonetic category segment in the utterance. The sliding window approach, used in Section 4.3.1 for high-quality speech, was used here. The salient aspects of the approach are summarized below.

Let L be the number of languages being classified. Features from a moving window of N segments were presented to the network at a time. The network produced L language activation scores for each window of N segments. There was an increment of $I(< N)$ segments between successive windows. The L language activation scores were accumulated as the window progressed through the utterance. When the end of the utterance was reached, the language with the maximum accumulated activation score was taken as the system response.

5.4.2.1 Feature Computation

For each segment in a window, the following feature measurements were made:

- the segment duration
- a vector of length seven, representing the broad phonetic label. There were two experimental conditions depending on the composition of this vector:
  - It was either a binary-valued vector, i.e., the element corresponding to the broad phonetic category label of the segment was set to 1.0 and the rest set to -1.0, or,
  - it contained the averaged frame-based activation scores from the segmenter network for each of the seven broad phonetic categories. The averaging was done over all frames within the segment.

Thus, there were a total of 8 features from each segment. Given that on the average, there were about 120 segments per utterance and 271 utterances per language in the training set, a large number of feature vectors were generated per utterance ($960 = 8 \times 120$), and per language ($260160 = 960 \times 271$), making it computationally infeasible to train a single network using vectors from all ten languages at once. Therefore, nine separate pairwise networks, with English as the anchor language, were trained and evaluated.

5.4.2.2 Network Architecture and Training

The number of input neurons was determined by the value of N, size of the moving window. The number of feature vectors (or windows) obtained per utterance was determined by the value of I, the window increment. The smaller the value of I, the greater the number of windows in an utterance and hence the greater the number of feature vectors. Parametric experiments were conducted on just the English-Japanese network to determine the optimal values for N and I. For these experiments, the label vector was binary-valued. The optimal language identification performance was obtained for $N = 15$ and $I = 1$ (single-segment increment). This configuration was then used to determine the optimal composition of the label vector. It was found that the averaged segmenter network scores outperformed the binary values. The results of these parametric experiments are described in the next sub-section.

Thus, for the language-pair experiments, with $N = 15$ and $I = 1$, the networks had 120 ($= 15 \times 8$) input neurons, either 24 or 32 hidden neurons, and 2 output neurons, with the number of hidden neurons being determined experimentally for each language pair.

5.4.2.3 Normalization

The input features were normalized to lie between 1.0 and -1.0 using a generalized form of Equation 4.1. It is worth specifying this generalized normalization formula, since it was used to normalize almost all input features described in this chapter.

$$normF_i = 2.0 \times \left(\frac{F_i - minF_i}{maxF_i - minF_i}\right) - 1.0 \tag{5.1}$$

where F_i is the i^{th} feature under consideration, $minF_i$ and $maxF_i$ are the minimum and maximum values of F_i, and $normF_i$ is the normalized value of F_i. $minF_i$ and $maxF_i$ are obtained from the 5^{th} and 95^{th} percentiles of a histogram of F_i generated on the training data across all languages. Normalizing the features in this fashion maximizes the differences in the distributions (if any) for input to the network.

The segment duration feature was normalized with respect to the minimum and maximum duration values for that broad phonetic category over all languages. The averaged frame-based scores were normalized with respect to the minimum and maximum values within each label vector.

5.4.2.4 Language Identification Performance

Parametric Experiments with English and Japanese. With the binary-valued label vectors, the best network performance was 62.3%, with $N = 15$ and $I = 1$. The corresponding language identification score, i.e., the percentage of whole utterances correctly classified, was 75.9%. With the averaged frame-based scores in the label vector, the language identification score improved to 83.2%, a reduction in the error-rate of about 30%. This is to be expected, since for a given segment, the averaged segmenter activation scores impose a seven-tier ranking on the label values, providing more discriminant

information to the network than is possible with the binary-valued vector, which does not differentiate between six of the seven values. The results of the English-Japanese parametric experiments are shown in Table 5.3. "BV" and "SS" refer to the binary-valued vector and the segmenter scores, respectively. I_{net} and I_{score} refer to the segment increments used for testing the network and evaluating the language identification score respectively. The *Network %* column shows the percentage of segment windows correctly classified. The *Score %* column shows the language identification score. Training the network with $I_{net} > 1$, and scoring with $I_{score} = 1$ (indicated by scores with asterisks) did have the advantage of reducing the training file sizes and thus the training time, but produced only very slight improvements in the identification scores (68.6% to 70.0% for $I_{net} = 5$, and 78.2% to 80.5% for $I_{net} = 2$). Training and testing with $I_{net} = I_{score} = 1$ provided superior performance in all cases.

Table 5.3: Baseline Experiments using English and Japanese

Label Vector	N	I_{net}	I_{score}	*Network %*	*Score %*
BV	15	1	1	62.3	75.9
BV	15	5	5	60.8	68.6
BV	15	5	1	60.8	70.0*
BV	22	1	1	63.0	72.6
BV	30	1	1	63.8	70.9
SS	15	1	1	66.8	83.2
SS	15	2	2	64.6	78.2
SS	15	2	1	64.6	80.5*

Other Language-Pairs. The pairwise classification experiment with $N = 15$, $I_{net} = 1$ and $I_{score} = 1$ was repeated with the remaining eight languages, with English as the anchor language. Table 5.4 displays the results of these experiments. The English-Japanese result is included for the sake of completeness. The language pairs are denoted by pairs of two-letter codes made up of the first two letters of the language names. The columns labeled $L1$ and $L2$ show the identification scores of the first and second languages in each pair. With the exception of Farsi, Korean and Vietnamese, the performance of

English is always lower than that of the second language in the pair, indicating that English was more likely to get confused with the second language in these pairs than vice-versa. This effect was most pronounced for English and German, both Germanic languages (42.3% and 72.0% respectively).

To summarize, with just the segment label and duration information from utterances that are 13.4 seconds long on the average, it is possible to distinguish between two languages with an accuracy ranging from 57.0% (English-German) to 85.3% (English-Tamil). The median score was 74.0%.

Table 5.4: Baseline Experiments: Language Pairs with English as Anchor

Language Pair	*Network %*	*L1 Score (%)*	*L2 Score (%)*	*Overall Score (%)*
EN-FA	63.1	69.4	69.2	69.3
EN-FR	63.1	69.4	74.6	72.0
EN-GE	55.7	42.3	72.0	57.6
EN-JA	66.8	74.8	91.8	83.2
EN-KO	60.3	68.5	62.0	65.3
EN-MA	63.1	72.1	75.9	74.0
EN-SP	66.3	68.5	82.0	75.2
EN-TA	75.2	82.9	87.9	85.3
EN-VI	66.8	81.1	70.2	75.8

5.4.3 Experiments with Spectral Features

Given that static neural networks require equal-length feature vectors, how does one present spectral features, derived from utterances of varying length, as input? There are two levels of granularity: frame-level and segment-level spectral features. With frame-level spectral features, the following questions need to be addressed: How does one pick the frames from which the spectral information is sampled? What number of frames is considered "optimal?" If an utterance is shorter than the chosen fixed number, the feature vector will have to be padded with zeros. Zero-valued features might be misleading to the network. If the utterance is very long, then the spectral sampling might be too coarse. With segment-level spectral features, the usual method is to average the spectral

coefficients from different subdivisions of the segment. The main concern here is the amount of smearing caused by the averaging. The greater the number of subdivisions of the segment, the greater the spectral variation preserved across the segment, and the larger the dimension of the input feature vectors.

Two experiments were conducted using just PLP-based spectral features to identify English and Japanese. The features were designed to be independent of the utterance duration by using spectral information at the frame and segment levels.

5.4.3.1 Spectral Experiment I: Frame-based PLP Scores

A baseline approach to using spectral features is to classify each 3 ms frame as one of the two languages using a neural network, and accumulate the network output activations across all frames of the utterance for each language. The language with the maximum accumulated activation score is the winner. This approach does not use broad phonetic category information.

Feature Computation. The spectral features consisted of 56 averaged PLP coefficients taken from 7 regions spanning a 171 ms window centered on the frame to be classified. The sampling intervals are shown in Figure 5.2. The solid boxes indicate 9 ms and 15 ms intervals over which PLP coefficients are averaged. Dashed boxes indicate 18 ms intervals that are skipped. The objective was to provide substantial contextual information about the chosen frame to the network.

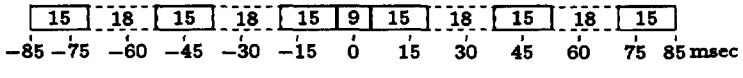

Figure 5.2: Spectral Experiment I: Sampling intervals for the PLP features.

Frame Sampling and Network Training. Since it was not computationally feasible to train a network using features from all frames in each utterance, frames were sampled at fixed intervals from each utterance. Two intervals were evaluated, 24 ms (8 frames)

and 45 ms (15 frames). The smaller the interval, the greater the number of feature vectors generated. Resource constraints prevented experimentation with smaller intervals. Larger intervals were deemed useless based on the reduction in performance with the 45 ms interval.

Results. Several network configurations were evaluated. Table 5.5 summarizes the frame-by-frame classification performance of the different networks. The best network performance, 59.0%, was obtained from the 56-48-2 network trained on data with a 24 ms sampling interval. The results indicate that all networks performed at approximately the same level of accuracy. The test utterances were scored on the two best networks. The language identification score corresponding to the 56-42-2 network was 70.0% and that of the 56-48-2 network was 69.2%. The results of this spectral experiment provide a baseline for English-Japanese identification performance. Using just spectral information and no segmentation, English and Japanese were identified with 70% accuracy. In the following sections, performance on this task using more sophisticated features is examined.

Table 5.5: Spectral Experiment I: Identification Results

Sampling Interval (ms)	*Configuration*	Frame Acc. (%)	*Id. Score (%)*
24	56-32-2	58.0	-
24	56-36-2	58.8	-
24	56-42-2	59.0	70.0
24	56-48-2	59.0	69.2
45	56-28-2	57.7	-
45	56-32-2	58.2	-
45	56-36-2	58.5	-

5.4.3.2 Spectral Experiment II: Averaged PLP Coefficients in Each Segment

Feature Computation. PLP coefficients were averaged across all frames within each segment so that a single vector of 8 spectral features was obtained from each segment. The sliding window approach described in Section 4.3.1 for label and duration was then

used with $N = 15$ and $I = 1$, so that each input feature vector had 120 ($= 15 \times 8$) spectral features.

Results. The best network configuration, 120 ($= 15 \times 8$) inputs, 18 hidden units and 2 output units, performed with an accuracy of 58.8%. The corresponding language identification score was 65.0%. This comparatively low result is to be expected given that the averaging smeared over a lot of spectral variation within each broad phonetic segment.

5.4.4 Adding Spectral Information to Baseline Features

Given the encouraging result of the label + duration baseline experiment (with averaged segmenter network scores) and the two baseline spectral experiments, the next logical step was to examine the effect of adding each of the two spectral feature sets to the 8 label and duration features computed for each segment. For the sake of efficiency, the combination experiments were conducted with English and Japanese only, with the objective of extending it to other language pairs only if significant performance improvements were obtained.

5.4.4.1 Combination Experiment I: Label + Duration + Frame-based PLP scores

Procedure. The PLP features for each segment were obtained as follows: Frame-based PLP network activation scores (from the PLP network in Spectral Experiment I) were accumulated for each language within the segment. The accumulated score was then divided by the segment duration (in frames) to get the averaged PLP activation score per frame for that language. The two averaged scores (for English and Japanese) were added to the eight label and duration features to obtain 10 features per segment. The sliding window approach was then used with $N = 15$ and $I = 1$. Each input vector to the network contained 150 ($= 15 \times 10$) features.

Results. The best network performance, 72.7%, was obtained with a 150-42-2 network. When scored on the development set utterances, however, the language identification performance was only 74.4%. Comparing this to the 83.2% obtained by the label and duration features alone (Section 5.4.2.4) and the 70.0% obtained with the frame-based PLP features alone (Section 5.4.3.1) it appears that the two PLP network activation scores per segment actually hurt identification performance.

5.4.4.2 Combination Experiment II: Label + Duration + Averaged PLP in Segment

Procedure. For each segment, PLP coefficients averaged across the entire segment (8 values) were added to the 8 label and duration features for a total of 16 features per segment. The sliding window approach with $N = 15$ and $I = 1$ was used. Each input vector to the network contained 240 ($= 15 \times 16$) features.

Results. The best network performance, 67.7%, was obtained using a 240-75-2 network. The language identification score on the development set utterances was 80.5%. This result represents a 23.8% reduction in error-rate compared to the score obtained by adding the frame-based PLP scores to the label and duration features (74.4%) and a substantial 44.3% reduction in error-rate compared to using the PLP coefficients alone (65.0%).

5.4.4.3 Discussion

The results of the experiments with the baseline and spectral features on the English-Japanese task are summarized in Table 5.6. The averaged PLP coefficients perform much better in conjunction with the label and duration features (80.5%) than in isolation (65.0%). Adding the PLP coefficients to the label and duration features actually lowered performance from 83.2% to 80.5%, an increase in the error-rate of 16.1%.

On the other hand, the frame-based PLP features (70.0%) performed better than the averaged PLP coefficients (65.0%) in isolation. When added to the label and duration

Table 5.6: Summary of Baseline and Spectral Experiments on English and Japanese

Feature	*Accuracy (%)*
Dur. + Label	83.2
Frame-based PLP scores	70.0
Avge. PLP in each segment	65.0
Dur. + Label + Frame-based PLP	74.4
Dur. + Label + Avge. PLP	80.5

features however, the averaged frame-based PLP network scores caused a steeper drop in performance (83.2% to 74.4%) than the averaged PLP coefficients (83.2% to 80.5%). In either case, addition of PLP-based spectral features does not help overall language identification performance. The label and duration features still produce the best language identification score of 83.2%. Based on these experiments, it was not considered worthwhile to investigate PLP-based spectral features further in this dissertation.

5.4.5 Boxplots

Boxplots provide a very powerful method for visualizing the rough distributional shape of two or more data distributions. They were extensively used in the exploratory analyses of the features described in this chapter.

A boxplot [Sta91] is a graphical representation showing the center and spread of a data distribution, along with a display of unusually deviant data points, called *outliers*. Figure 5.3 shows an example of a boxplot. The horizontal line in the interior of the box is located at the median of the data. This estimates the center of the distribution. The height of the box is equal to the interquartile distance, or IQD, which is the difference between the third and first quartiles of the data. The IQD indicates the spread or width of the distribution for the data. It encompasses the middle half of the data points in the distribution. The *whiskers* (dotted lines extending from the top and bottom of the box) extend to the extremes of the data or $1.5 \times IQD$ from the center, whichever is less. For data having a Gaussian distribution, approximately 99.3% of the data falls inside the

whiskers. Data points which fall outside the whiskers may be outliers, and so they are indicated by horizontal lines.

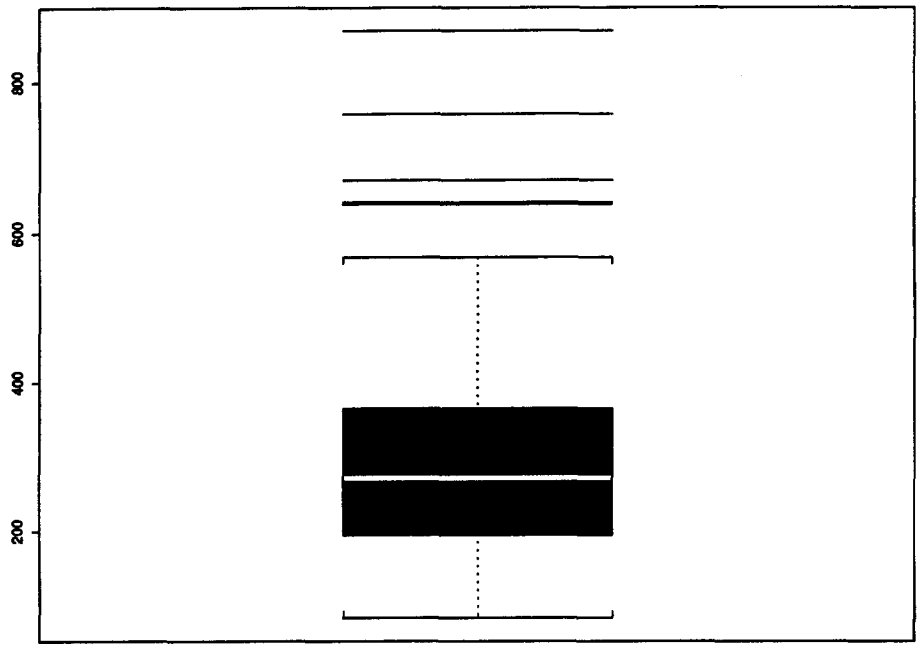

Figure 5.3: Example of a Boxplot

For the feature analyses described in this chapter, boxplots were generated both for the un-normalized feature values and for the normalized features. The normalized-feature boxplots served to display the differences in the language distributions exactly as presented to the network. This proved to be a useful aid in the feature selection process. Features for which the normalized boxplot did not show sufficient language discrimination (i.e., separation along the ordinate) were discarded.

5.4.6 Segment-triples

Informal evidence suggests that languages differ in the frequency and patterns of co-occurrence of phonemes. It would be interesting to see if such differences are retained at

the broad phonetic level. Features were examined in the order of increasing granularity: features based on triples of segments were examined first, followed by those based on segment-pairs, followed by features based on individual segments.

There are 63 legal segment-triples of the seven broad phonetic categories. These are listed in Table 5.7 along with their 3-letter codes. The following features were computed

Table 5.7: List of 63 Legal Segment-triples

Segment-triple	*Code*	*Segment-triple*	*Code*	*Segment-triple*	*Code*
CLOS FRIC CLOS	CFC	INVS VOC CLOS	IVC	STOP CLOS STOP	SCS
CLOS FRIC PRVS	CFP	INVS VOC FRIC	IVF	STOP CLOS VOC	SCV
CLOS FRIC VOC	CFV	INVS VOC INVS	IVI	STOP FRIC CLOS	SFC
CLOS PRVS VOC	CPV	INVS VOC POVS	IVP	STOP FRIC PRVS	SFP
CLOS STOP CLOS	CSC	POVS CLOS FRIC	PCF	STOP FRIC VOC	SFV
CLOS STOP FRIC	CSF	POVS CLOS PRVS	PCP	STOP PRVS VOC	SPV
CLOS STOP PRVS	CSP	POVS CLOS STOP	PCS	STOP VOC CLOS	SVC
CLOS STOP VOC	CSV	POVS CLOS VOC	PCV	STOP VOC FRIC	SVF
CLOS VOC CLOS	CVC	POVS FRIC CLOS	PFC	STOP VOC INVS	SVI
CLOS VOC FRIC	CVF	POVS FRIC PRVS	PFP	STOP VOC POVS	SVP
CLOS VOC INVS	CVI	POVS FRIC VOC	PFV	VOC CLOS FRIC	VCF
CLOS VOC POVS	CVP	POVS STOP CLOS	PSC	VOC CLOS PRVS	VCP
FRIC CLOS FRIC	FCF	POVS STOP FRIC	PSF	VOC CLOS STOP	VCS
FRIC CLOS PRVS	FCP	POVS STOP PRVS	PSP	VOC CLOS VOC	VCV
FRIC CLOS STOP	FCS	POVS STOP VOC	PSV	VOC FRIC CLOS	VFC
FRIC CLOS VOC	FCV	PRVS VOC CLOS	PVC	VOC FRIC PRVS	VFP
FRIC PRVS VOC	FPV	PRVS VOC FRIC	PVF	VOC FRIC VOC	VFV
FRIC VOC CLOS	FVC	PRVS VOC INVS	PVI	VOC INVS VOC	VIV
FRIC VOC FRIC	FVF	PRVS VOC POVS	PVP	VOC POVS CLOS	VPC
FRIC VOC INVS	FVI	STOP CLOS FRIC	SCF	VOC POVS FRIC	VPF
FRIC VOC POVS	FVP	STOP CLOS PRVS	SCP	VOC POVS STOP	VPS

for segment-triples:

- Segment-triple Frequency (STF): number of occurrences of each segment-triple per second of speech, and

- Segment-triple Ratio (STR): ratio of the number of occurrences of each segment-triple to the total number of segments in the utterance

The STR features were considered essential since the STF features alone might present a misleading picture, given that speech rates (number of segments per second) vary widely across speakers.

An analysis of variance using the one-way layout model [BHH78, Sta91] was conducted to determine the usefulness of the 126 segment-triple features in distinguishing between the 10 languages. Of the 63 segment-triples, five produced p-values greater than 0.01. These were discarded. The five useless segment-triples and their p-values are given below:

- POVS-STOP-CLOS 0.011
- POVS-STOP-FRIC 0.237
- POVS-STOP-PRVS 0.166
- STOP-CLOS-VOC 0.573
- STOP-FRIC-PRVS 0.319

Therefore, 116 features from 58 segment-triples were used in all the experiments. Performance of these features on the ten-language task and the two-language task (English-Japanese only) were examined.

5.4.6.1 Ten-language Task

Several networks were trained on the 116 STF and STR features. The best performance, 40.9%, was obtained with a 116-50-10 network. To determine if all 116 features were contributing to the classification performance, two techniques were used to reduce the feature set:

- A principal component analysis was performed on the 116 features to determine the components of maximum variance in the feature set, and

- "Useful" features were selected by visual inspection of the boxplots of all 116 features.

Principal Component Analysis (PCA). This technique has been shown to be useful in reducing the dimensionality of the feature space for speech recognition tasks with negligible loss in performance [LRH90]. There is no rigorous method of determining the optimal number of principal components of a given set of vectors. A good heuristic, however, is to generate a plot of the logarithm of the eigenvalues of the features. The number of eigenvalues corresponding to the "knee" of this curve is then taken as the "optimal" number of principal components. To double-check this procedure, the following ratio is computed:

$$V_k = \sum_{m=1}^{k} \lambda_m / \sum_{m=1}^{N} \lambda_m, 1 < k < N \tag{5.2}$$

where

- V_k is the fraction of the variance retained by the first k eigenvalues
- N is the total number of eigenvalues (= original number of features),
- k is the number of eigenvalues corresponding to the knee of the curve, and
- λ_m is the m^{th} eigenvalue.

If $V_k \geq 0.90$, then the first k principal components of the feature set are chosen, as they account for at least 90% of the variance. Figure 5.4 shows a plot of the logarithm of the eigenvalues of the 116 segment-triple features. The knee of this curve is at 40 and $V_{40} = 0.971$, indicating that the first 40 principal components of the 116 segment-triple features account for 97.1% of the variance. The first 40 principal components of the segment-triple features were computed, normalized to lie between -1.0 and 1.0, and given as input to a neural network classifier. The best performance, 40.5%, was obtained with a 40-27-10 network. This represents a negligible increase in the error-rate compared to the performance on the full set of 116 features with a 65% reduction in the dimension of the input features.

Visual Inspection. Appendix C.1 shows the boxplots of all the 116 segment-triple features. There are two boxplots per feature, one with the un-normalized features, and the other with the normalized features.

The criterion used to declare a feature useful was simple: the ten boxes should show some degree of separation along the ordinate. For example, if all boxes had their medians (the white horizontal bands) aligned, then that feature is clearly not discriminant enough to be included in the set of useful features. It is to be stressed that this selection process was based on informal visual evidence rather than quantitative measures.

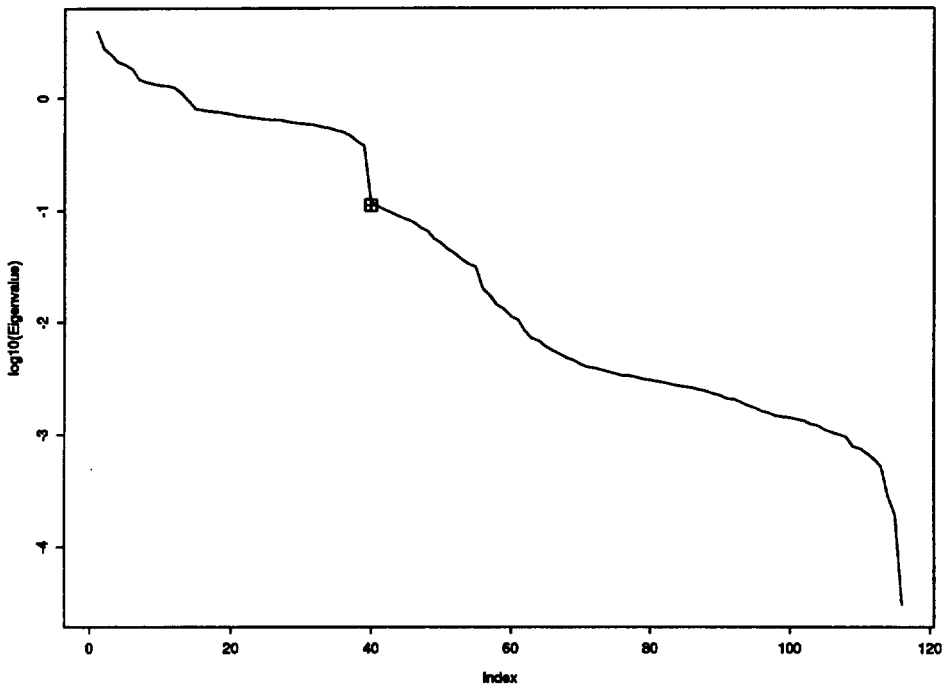

Figure 5.4: Logarithm of the Eigenvalues of the 116 STF and STR Features

Using this procedure, 27 STF and 30 STR features were selected for a total of 57 features out of the original 116. The selected trigrams are listed in Table 5.8. A 57-15-10 network trained on these 57 segment-triple features performed at an accuracy of 32.6%. This represents a 14.0% increase in the error-rate (and a 50.8% reduction in the dimension of the input features) compared to the original 116-feature result. Apparently,

our procedure of visual selection was too restrictive and removed useful features!

Table 5.8: List of 57 Useful Segment-triple Features

STF	*STR*
CFV CSV FCS FVC FVP IVI PCS PFV SVC SVF SVI VCF VCS VFV VIV VPF VPC FVI CFC IVP IVC CSC PCF SVP PVI IVF FVF	CFV CSV FCS FVC FVP IVI PCS PFV SVC SVF SVI VCF VCS VFV VIV VPF CPV VPC FVI CFC IVP IVC CSC PVP PCF SVP PVI VFC IVF FVF

5.4.6.2 Two-language Task

Rather than examine all 45 language pairs or even the nine $EN-L'$ pairs, it was decided to concentrate only on the English-Japanese classification task for the segment-triple experiments.

Visual inspection of the segmentation output for English and Japanese utterances indicated that Japanese utterances had more consecutive occurrences of the following segment-triples: (i) CLOS-STOP-VOC (CSV), (ii) VOC-CLOS-STOP (VCS), and (iii) STOP-VOC-CLOS (SVC). The following features were examined for each of the three segment-triples:

- LSQ - longest sequence in an utterance
- LSQF - longest sequence per second of speech
- LSQ10 - longest sequence in a moving 10-segment window
- #SQGT1 - number of sequences greater than one, per second of speech
- #SQGT2 - number of sequences greater than two, per second of speech
- #SQGT3 - number of sequences greater than three, per second of speech

where "sequence" referred to two or more consecutive occurrences of the segment-triple.

These features were examined in isolation and in combination. Table 5.9 summarizes the results of these segment-triple experiments. From the table it is clear that the #SQGT3 is not useful, given its almost-chance classification performance for all three segment-triples. Also, #SQGT1 appears to be the most useful single feature for all three triples. The best performance, 67.4%, was obtained using the #SQGT1 feature for the triple CSV and #SQGT1 feature for the triple VCS together. With all 18 features, the performance was 66.5%, reflecting the uselessness of the other 16 features.

5.4.6.3 Summary of Segment-triple Experiments

The full set of 116 segment-triple frequency and ratio features provided 40.9% identification accuracy on the ten-language task using utterances that were 13.4 seconds long on the average. The first 40 principal components of these features performed at 40.5%, a negligible drop in performance coupled with a 65% reduction in the dimension of the input features. On the English-Japanese task, #SQGT1 was identified as the most useful feature, while #SQGT1 for the triples CSV and VCS provided the best identification performance (67.4%).

Table 5.9: Segment-triple Experiments on English and Japanese

	Feature					
Segment-triple	#SQGT1	#SQGT2	#SQGT3	LSQ	LSQF	LSQ10
CLOS-STOP-VOC	66.1	60.8	52.9	59.0	58.6	59.5
VOC-CLOS-STOP	63.0	58.6	53.3	58.6	62.6	59.0
STOP-VOC-CLOS	62.1	57.3	51.5	62.1	61.7	62.1

5.4.7 Segment-pairs

There are 20 legal segment-pairs of the seven broad phonetic categories, VOC, FRIC, CLOS, STOP, PRVS, INVS and POVS. These 20 pairs are listed along with their two-letter codes in Table 5.10. Four feature sets based on segment-pairs were examined:

- Segment-pair Frequency (SPF): number of occurrences of each segment-pair per second of speech, and
- Segment-pair Ratio (SPR): ratio of the number of occurrences of each segment-pair to the total number of segments in the utterance
- Segment-pair Median Duration (SPMD): median duration of each segment-pair in an utterance
- Segment-pair Duration Ratio (SPDR): ratio of the total duration of all occurrences of a segment-pair in an utterance to the total utterance duration

As with segment-triples, the SPR features were considered essential since the SPF features alone might present a misleading picture, given that speech rates (number of segments per second) vary widely across speakers. The SPMD features sought to extract the differences (if any) in the median durations of the segment-pairs. The SPDR features determined the fraction of an utterance's duration that was accounted for by specific segment-pairs.

Table 5.10: List of 20 Legal Segment-pairs

Segment-pair	*Code*	*Segment-pair*	*Code*
VOC FRIC	VF	CLOS PRVS	CP
VOC CLOS	VC	STOP VOC	SV
VOC INVS	VI	STOP FRIC	SF
VOC POVS	VP	STOP CLOS	SC
FRIC VOC	FV	STOP PRVS	SP
FRIC CLOS	FC	PRVS VOC	PV
FRIC PRVS	FP	INVS VOC	IV
CLOS VOC	CV	POVS FRIC	PF
CLOS FRIC	CF	POVS CLOS	PC
CLOS STOP	CS	POVS STOP	PS

These 80 features above were motivated by informal knowledge that languages differed in the characteristics of pairs of broad phonetic categories.

An analysis of variance using the one-way layout model was conducted to determine the usefulness of the 80 features in distinguishing between the 10 languages. All 80 features produced p-values less than 0.01 indicating some degree of discriminative power.

Language identification experiments were done with segment-pair features on the ten-language task and a two-language task, the latter being confined to English and Japanese only. The ten-language task was chosen to provide insights into the effectiveness of segment-pair characteristics in distinguishing between this varied mix of languages. The two-language task was inspired by visual inspection of boxplots generated for English and Japanese that showed marked separation in feature values for certain segment-pairs.

5.4.7.1 Ten-language Task

Networks were trained with all 80 features as input. The performance accuracy was 43.5% on the development set. Scatter-plots were generated between all pairwise combinations of the four feature sets for a few segment pairs. A scatter-plot is useful in indicating correlations between feature pairs. If all the data points in a scatter-plot align themselves on or close to a straight line (of any slope), then one can safely assume that the two features are highly correlated. Figure 5.5 and Figure 5.6 display the plots for the segment-pairs VOC-FRIC and INVS-VOC. The following consistent trends emerged from the plots:

- SPF, SPR and SPDR were highly correlated.
- SPMD and SPDR were *not* correlated.
- SPMD is not correlated with either SPF or SPR.

The high correlation between the feature sets suggested that many of the 80 features might be redundant. Principal component analysis and visual inspection were again used to reduce the redundancy in the feature sets.

PCA. Figure 5.7 displays the eigenvalue curve for the 80 features. The knee of the curve is at 36, and $V_{36} = 0.984$ (from equation 5.2). Therefore, the first 36 principal

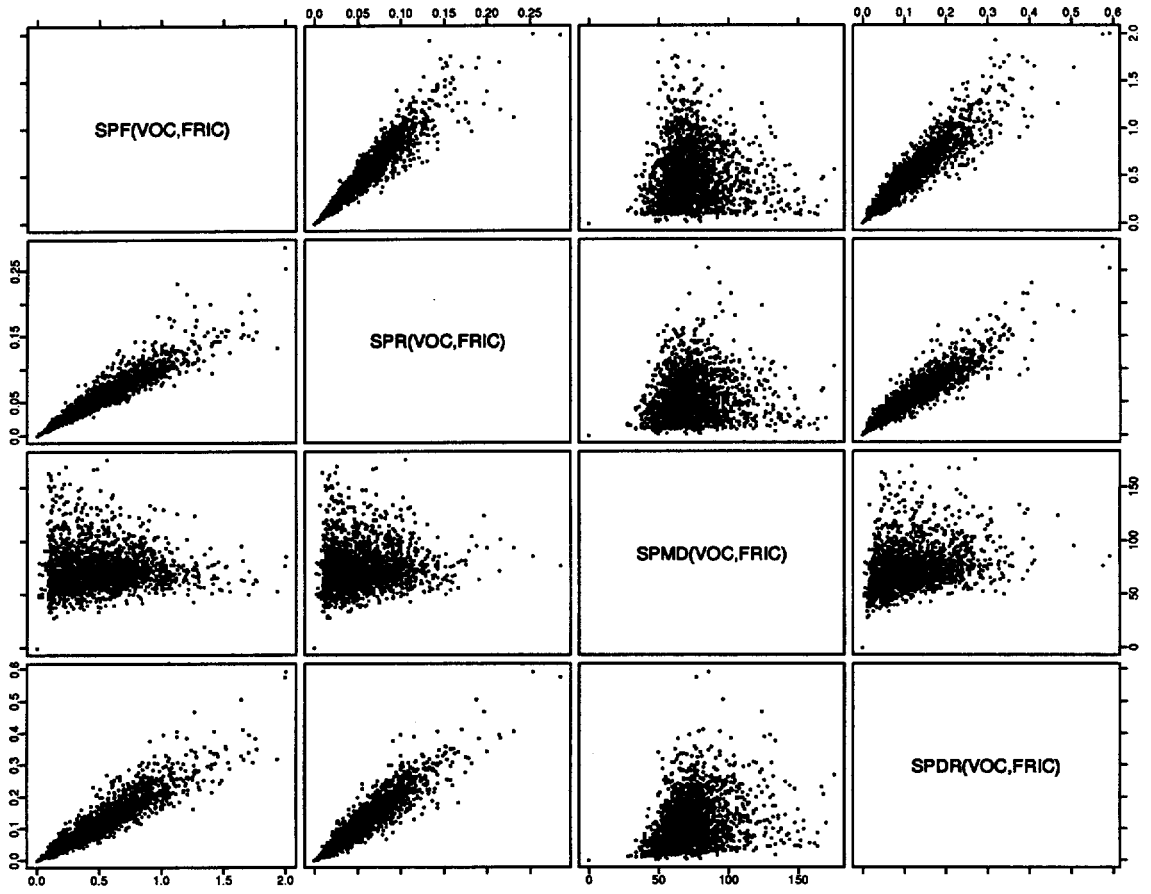

Figure 5.5: Scatterplots of SPF, SPR, SPMD, and SPDR for VOC-FRIC

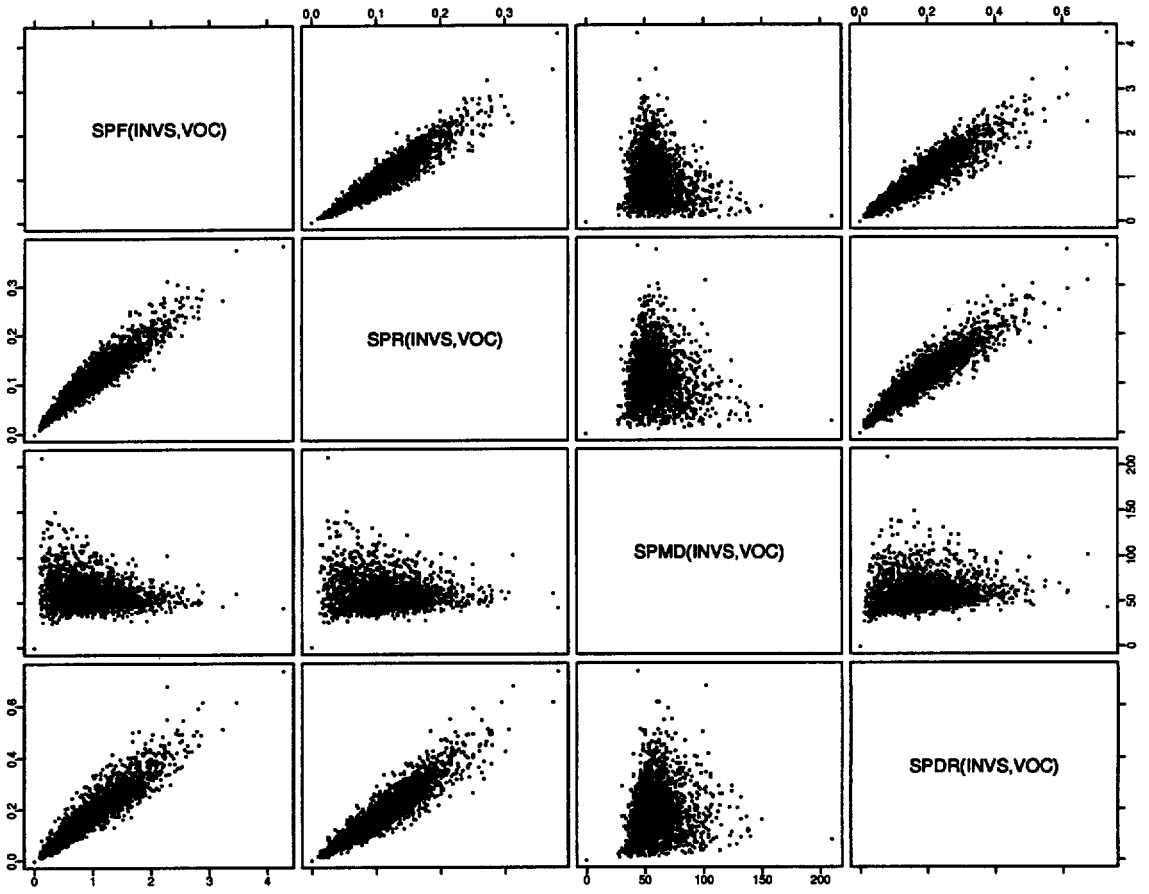

Figure 5.6: Scatterplots of SPF, SPR, SPMD, and SPDR for INVS-VOC

components of the 80 segment-pair features were computed, normalized to lie between -1.0 and 1.0, and fed into a neural network. The best classification performance, 42.1%, was obtained with a 36-38-10 network trained on these 36 features. This represents a 2.5% increase in the error-rate compared to the 80-feature network, with a 55% reduction in the dimension of the input features.

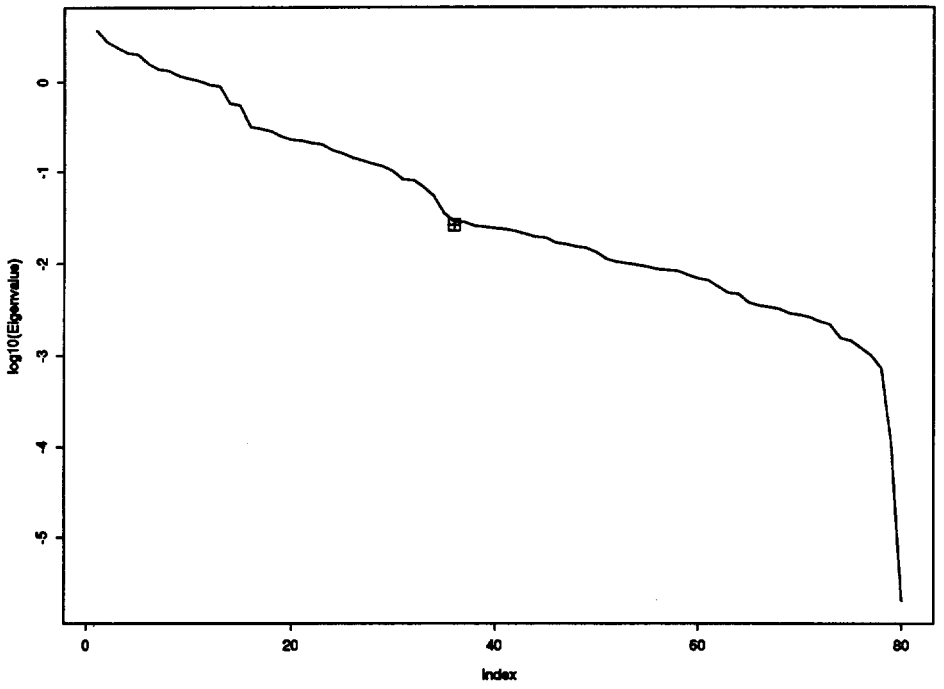

Figure 5.7: Logarithm of the Eigenvalues of the 80 SPF and SPR Features

Visual Inspection. Figures C.43 through C.70 in Appendix C.2 shows the boxplots of all the 80 segment-pair features. There are two boxplots per feature, one with the un-normalized features, and the other with the normalized features.

By visual inspection of the boxplots, 12 SPF, 12 SPR, 6 SPMD and 10 SPDR features were selected for a total of 40 features out of the original 80. The selected features are listed in Table 5.11. A 40-30-10 network trained on these 40 segment-pair features performed at an accuracy of 43.5%. This is comparable to the performance of the 36

principal components network, and identical to that obtained with the original 80-feature network. Thus, our visual inspection proved successful in this case.

Table 5.11: List of 40 Useful Segment-pair Features

SPF	*SPR*	*SPMD*	SPDR
VF VC VF VC VI VP FV FC CF CS SV IV PF PC	VF VC VF VC VI VP FV FC CF CS SV IV PF PC	VF VI VP SV IV PF	VF VC VI VP FV FC CF IV PF PC

5.4.7.2 Two-language Task

Visual inspection of the boxplots for the four feature sets (SPF, SPR, SPMD and SPDR) with respect to English and Japanese revealed that there was some separation in the two distributions for SPF, SPR and SPDR for the following segment-pairs:

1. POVS-FRIC
2. CLOS-STOP
3. VOC-CLOS
4. VOC-POVS

Of these, VOC-POVS clearly had the maximum difference in distributions. Interestingly, there were no distributional differences for the SPMD feature set indicating that English and Japanese differed only in the frequency and ratios of segment-pairs rather than in their median durations.

To determine the contributions of these 12 features to the English-Japanese classification task, a series of neural network experiments were conducted using these features in isolation. Table 5.12 summarizes the results of these experiments. The informal evidence from the boxplots, namely the wide difference between the EN and JA distributions for

all three VOC-POVS features, was borne out by the experimental results. The SPR feature for the segment-pair VOC-POVS produced the best single-feature result of 74.5%. With all 12 features, performance improves to just 75.8% with a 12-10-2 network, indicating the dominance of the VOC-POVS features for English-Japanese classification.

Table 5.12: Segment-pair Experiments on English and Japanese

	Single Feature Result (%)		
Segment-pair	SPF	SPR	SPDR
POVS-FRIC	59.0	57.1	57.7
CLOS-STOP	63.4	59.5	54.6
VOC-CLOS	62.6	62.6	65.2
VOC-POVS	71.8	74.5	73.6

5.4.7.3 Summary of Segment-pair Experiments

Language identification experiments were performed on a set of 80 frequency, ratio, duration ratio and median duration features derived from 20 legal segment-pairs of the seven broad phonetic categories. With all 80 features, the accuracy was 43.5%. Scatter plots of the features revealed high correlation between the SPF, SPR and SPDR features. The first 36 principal components performed at 42.1%, while 40 features derived through visual inspection of the boxplots performed at 43.5%—identical to the performance with the original 80 features. On the English-Japanese task, the ratio of occurrence of VOC-POVS proved to be the most useful feature, performing at an accuracy of 74.5%.

5.4.8 Pitch-based Features

5.4.8.1 Pitch Tracking

The pitch tracker used in this research was developed by Daniel Burnett [Bur92], and trained on multi-lingual data from the OGI Ten-language Telephone Speech Corpus.

A neural network-based voicing detector used frame-based PLP features to determine voiced and un-voiced regions of an utterance. The pitch tracker computed pitch periods using the method of autocorrelation, on voiced portions of speech only (as determined by the voicing detector). The output of the pitch tracker consisted of median pitch period values, one for every 3 ms frame of the utterance. Unvoiced frames have a value of 0. The reciprocal of each non-zero period value provided an estimate of median pitch in Hz for that frame.

5.4.8.2 Experiments

Playback of utterances from English and tone languages such as Mandarin and Vietnamese, and visual inspection of median pitch contours suggested that languages might differ in the intra-segment and inter-segment variations in fundamental frequency. With this knowledge in mind, the following features were computed:

- Intra-segment pitch variation: Average of the standard deviations of the pitch within all sonorant segments—VOC, PRVS, INVS and POVS (4 values)
- Inter-segment pitch variation: Standard deviation of the average pitch in all sonorant segments (4 values)

Figures C.71 through C.73 in Appendix C.3 display the boxplots of these 8 pitch features.

The effect of just these 8 pitch features on the ten-language classification task was examined by training several networks. The best performance, 18.8%, was obtained with a 8-15-10 network, indicating very poor discriminant ability of the 8 pitch-based features.

5.4.9 Other Global Features

Apart from the segment-pair and segment-triple features, a number of features were computed based on individual segment statistics. These features were the result of exploratory analyses of the segment sequences. There were a total of 129 segment-based features in this set. They are grouped by type in the sections below. Boxplots of these

features are given in Figures C.74 through C.117 in Appendix C.4. The number of feature values in each class is indicated in parentheses.

5.4.9.1 Frequency of Occurrence (11 features)

It is well known that languages differ in the frequency of occurrence of certain sounds. The features described below sought to determine if these differences extended to broad phonetic categories as well. In the list below, an "obstruent" is either a STOP or a FRIC. An obstruent was considered voiced if more than half the segment was labeled as voiced by the voicing detector. The voiced obstruent features were included to accentuate differences between languages with relatively high voiced obstruent frequencies (English, French, German) and those with low voiced obstruent frequencies (Spanish and Tamil). These differences are apparent in the boxplot labeled "freq(VOBS)" in Figure C.76 in Appendix C.4.1.

- Frequency of occurrence of each of the seven broad phonetic categories (7 values)
- Frequency of occurrence of all segments (number of segments per second) (1 value)
- Frequency of occurrence of all sonorants (1 value)
- Frequency of occurrence of all obstruents (STOPs and FRICs) (1 value)
- Frequency of occurrence of voiced obstruents (STOPs and FRICs) (1 value)

5.4.9.2 Segment Ratios (94 features)

As in the case of segment-pairs and triples, mere frequency of occurrence of individual broad phonetic categories does not give the complete picture. Ratio of occurrence, i.e., number of occurrences of a particular category divided by the total number of segments, in conjunction with the frequency of occurrence, normalizes speech rate differences among speakers of the same language. There were a total of 47 segment occurrence ratios. Figures C.78 through C.92 in Appendix C.4 display the boxplots of the features.

- Ratio of number of occurrences of each of the seven broad phonetic categories to the total number of segments (7 values)
- Ratio of number of occurrences of each of the seven broad phonetic categories to the total number of sonorants (7 values)
- Ratio of number of occurrences of each of the seven broad phonetic categories to the total number of obstruents (7 values)
- Ratio of number of occurrences of sonorants to that of obstruents (1 value)
- Ratio of number of occurrences of VOC, CLOS, STOP, PRVS, INVS and POVS to that of FRIC. (6 values)
- Ratio of number of occurrences of VOC, STOP, PRVS, INVS and POVS to that of CLOS. (5 values)
- Ratio of number of occurrences of VOC, PRVS, INVS and POVS to that of STOP. (4 values)
- Ratio of number of occurrences of VOC, INVS and POVS to that of PRVS. (3 values)
- Ratio of number of occurrences of VOC and POVS to that of INVS. (2 values)
- Ratio of number of occurrences of VOC to that of POVS. (1 value)
- Ratio of number of sonorants to total number of segments (1 value)
- Ratio of number of obstruents to total number of segments (1 value)
- Ratio of number of voiced obstruents to total number of segments (1 value)
- Ratio of number of voiced obstruents to total number of obstruents (1 value)

Apart from speech rate differences among speakers, one needs to examine intrinsic speech rate differences across languages. To do this, it is necessary to examine segment

duration ratios in conjunction with segment *occurrence* ratios. Segment duration ratios help in determining the fraction of the utterance duration devoted to individual broad phonetic categories. The 47 duration ratio features can be obtained from the above list by replacing "number of occurrences" and "total number" with "duration" and "total duration" respectively. Figures C.93 through C.107 in Appendix C.4 display the boxplots for the duration ratio features.

5.4.9.3 Duration (24 features)

Apart from ratios of durations of the individual categories, the following duration features were examined.

- Inter-segment duration difference: defined as the absolute difference in durations between successive segments. This feature was motivated by the knowledge (obtained by visual inspection of segmenter output) that the broad phonetic segments of Japanese utterances were of almost equal length. This resulted in low values of this feature for Japanese, but high values for a language like Mandarin which was characterized by long VOC segments and much shorter STOP and FRIC segments. The features computed were the minimum, median, average, standard deviation and maximum of the inter-segment duration difference. To avoid using outliers, the minimum and maximum values were obtained by generating a histogram of the duration differences and choosing the 5^{th} and 95^{th} percentile values. (5 values)

- Vowel center distance: defined as the distance between the centers of successive vowels. This feature was designed with the consonant clusters of English in mind (CCVCC syllable structure). This feature distinguished English from Japanese and Tamil, both of which have a CV syllable structure. The features computed were the minimum, median, average, standard deviation and maximum vowel center distance. (5 values)

- Average duration of the seven broad phonetic categories (7 values)

- Standard deviation of the duration of the seven broad phonetic categories (7 values)

5.4.9.4 Ten-language Task

Several networks were trained with these 129 features. The best performance, 41.1%, was obtained with a 129-80-10 network. The eigenvalue curve for this feature set is shown in Figure 5.8. The knee was taken at 19, with $V_{19} = 0.919$. With just the first 19 principal components, the performance dropped to 37.0%, a 7.0% increase in the error-rate. However, with the first 40 principal components as features, $V_{40} = 0.977$ and the performance was 40.1%, a negligible increase of 1.7% in the error-rate with a corresponding 69.0% decrease in the input feature dimension. This is one example of a case where the eigenvalue curve does not provide an accurate estimate of the "optimal" number of principal components.

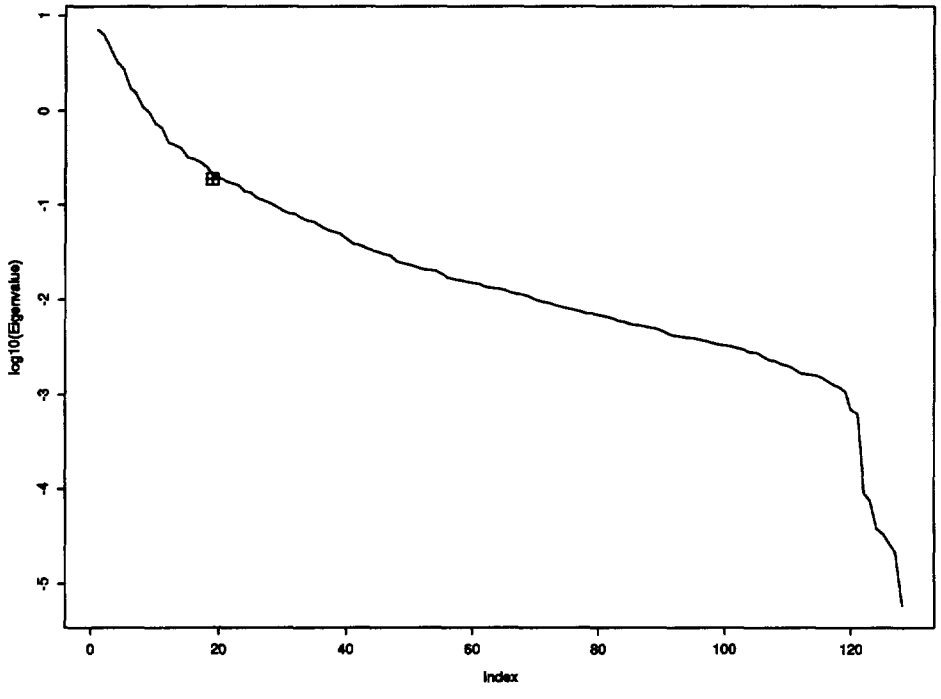

Figure 5.8: Logarithm of the Eigenvalues of the 129 Global Features

5.4.9.5 Two-language Task

Visual inspection of the boxplots indicated that certain features were useful for discrimination between certain language pairs. Specifically, the following observations were made:

1. Mandarin has a high frequency of occurrence of FRIC and a very low frequency of occurrence of INVS. Exactly the reverse is true for Tamil. This is evident in Figures C.74 and C.75. Also, given the approximately similar distributions of VOCs in Mandarin and Tamil, the ratio #VOC/#FRIC also seemed a good discriminant between the two languages.

2. English and Tamil also differ substantially in the frequency of occurrence of INVS and FRIC (see Figures C.74 and C.75).

3. Japanese has the lowest frequency of occurrence of POVS (Figure C.76). This feature was exploited in the segment-pair features described in Section 5.4.7.3.

The visual evidence described above was tested by training two-language networks using freq(INVS), freq(FRIC), #VOC/#FRIC in appropriate combinations. These experiments are summarized in Table 5.13. The results show that freq(INVS) is a better discriminator of Mandarin-Tamil than of English-Tamil. The reverse is true for freq(FRIC). Both features combined provide substantial improvements in performance for both language pairs. Of the single features, freq(FRIC) and #VOC/#FRIC provide the best classification performance on English-Tamil and Mandarin-Tamil, respectively. Addition of #VOC/#FRIC to the frequency features produces slight improvement in performance in the case of English-Tamil and none at all, in the case of Mandarin-Tamil.

Table 5.13: Some Pairwise Identification Results with Global Features

	Pairwise Id Score (%)	
Feature	*EN-TA*	*MA-TA*
freq(INVS)	68.0	72.1
freq(FRIC)	86.0	82.9
#VOC/#FRIC	81.6	87.4
freq(INVS)+freq(FRIC)	84.7	89.2
freq(INVS)+freq(FRIC)+#VOC/#FRIC	86.4	89.2

5.4.9.6 Summary of Experiments with Other Global Features

Language identification experiments were conducted with 129 features based on frequency of occurrence, segment occurrence ratios, segment duration ratios and inter-segment distances. These features were selected by a combination of linguistic knowledge, and visual inspection of the boxplots and segmenter output. On the ten-language task, the performance accuracy was 41.1%. It dropped to 40.1% with the first 40 principal components of the 129 features. On the two-language task, freq(INVS) and freq(FRIC) proved to be the most useful single features for the MA-TA and EN-TA tasks respectively. Further, the combination of just these two features provided 89.2% accuracy on the MA-TA task.

5.4.10 Combined Feature Set

In each of the sub-sections above, I have examined feature sets, both in their entirety and after reducing them using principal component analysis. Given that principal component analysis extracts features of maximum variance, it stands to reason that combining the principal components of each of the feature sets above would yield a compact feature set, relatively free of redundancies. Table 5.14 lists the components of the combined feature set and the corresponding numbers of principal components. The combined full-blown feature set consisted of 333 phonetic and prosodic features. The aggregate of the principal components of the four feature sets was 124. Given the small number

of pitch-based features, it was not considered essential to reduce these eight features further. The classification performance on this 124-feature set was then contrasted with that on the combined 333-feature set. In addition to the ten-language task, identification performance was also examined on language pairs and triples with English as an anchor.

To determine the effect of utterance duration on the identification accuracy, classifiers were evaluated on just the 50-second "story" utterances in the development set, in addition to the variable duration utterances (the entire development set). In the remainder of this chapter, the variable length utterances in the development set (average duration: 13.4 seconds) will be referred to as the *short* utterances, while the "story" utterances (average duration: 45.3 seconds) will be called the *long* utterances.

Table 5.14: Combined Feature Set and Principal Components

Feature Set	*Full-blown*	*Principal Components*
Segment-pairs	80	36
Segment-triples	116	40
Pitch	8	8
Global Features	129	40
Combined	333	124

5.4.10.1 Ten-language Task

With the 333 features, the best performance was obtained by a 333-60-10 network: 48.5%. With the 124 principal components, a 333-70-10 network performed at 47.9%, a negligible increase in the error-rate of 1.2% accompanied by a 62.8% reduction in the dimension of the input features. Efficiency considerations clearly point to the 124-feature network, since it performs as well as a network with almost 3 times as many features. However, it was decided to use the 333-feature set as the "best combined feature set". This decision was motivated by two considerations:

- Given that the objective is to determine the feature set that provides the highest classification performance on the development set, the 333-feature set is clearly the

logical choice, however small the improvement.

- At this stage in our research, efficiency and speed of classification are not as important as accuracy. There are several other areas that can be targeted for speed-up (e.g., performing the signal processing in hardware, pipelining the processing of the waveform) that would provide much more substantial improvements in efficiency than is achievable with a smaller network size.

On the long utterances, the best identification performance, 65.6%, was obtained by a 333-70-10 network. The confusion matrices for identification of the short utterances and long utterances are displayed in Figures 5.9 and 5.10 respectively. It is interesting to note that with Vietnamese, there is actually a very slight decrease in the accuracy when evaluated on the long utterances, and Korean shows only a slight improvement as the average duration of the test utterances increases from 13.4 seconds to 45.3 seconds. Farsi shows the largest improvement of all languages (30.6% to 73.7%), with a 62.1% reduction in the error-rate.

label	EN	JA	MA	TA	SP	KO	VI	FR	GE	FA	
EN	39	6	14	2	9	11	10	6	8	10	33.9%
JA	5	46	1	6	19	4	2	16	7	6	41.1%
MA	10	1	64	7	0	10	3	2	7	5	58.7%
TA	0	4	3	71	8	4	15	1	1	6	62.8%
SP	6	4	2	11	62	7	4	10	2	3	55.9%
KO	3	7	9	11	7	36	10	8	6	12	33.0%
VI	7	3	7	10	1	3	66	4	0	6	61.7%
FR	5	12	7	2	6	8	3	59	11	2	51.3%
GE	16	6	6	1	4	8	3	5	66	3	55.9%
FA	16	6	14	5	6	9	4	7	10	34	30.6%

Figure 5.9: Confusion Matrix for 10-language Classifier: Short Utt.

label	EN	JA	MA	TA	SP	KO	VI	FR	GE	FA	
EN	9	0	1	0	1	2	0	1	1	3	50.0%
JA	0	13	0	0	2	1	0	0	1	0	76.5%
MA	0	0	14	2	0	1	0	1	0	1	73.7%
TA	0	1	0	14	3	1	0	0	0	0	73.7%
SP	0	0	0	1	15	0	0	2	0	1	78.9%
KO	0	0	1	1	0	7	3	1	0	5	38.9%
VI	0	0	1	1	0	3	11	1	0	1	61.1%
FR	1	1	2	0	1	2	0	12	1	0	60.0%
GE	2	2	0	0	0	2	0	0	13	0	68.4%
FA	1	1	0	0	0	1	0	0	2	14	73.7%

Figure 5.10: Confusion Matrix for the 10-language Classifier: Long Utt.

5.4.10.2 Using Fewer Than Ten Languages

Language Combinations. With the feature development complete, tasks involving fewer than ten languages could now be examined using the combined feature set. Experiments were conducted on the following language combinations:

- Nine $English - L'$ networks, where L' is one of the remaining nine languages.
- Ten $L - Other$ networks, where L was one of the ten languages, and *Other* consisted of an equal number of utterances picked at random from each of the remaining nine languages.
- Nine $English - L' - Other$ networks, which examined three-way classification. In this case, *Other* consisted of an equal number of utterances picked at random from each of the remaining eight languages.
- A $English - Japanese - Mandarin - Tamil$ network, for the sake of comparison with the high-quality speech language identification results described in Section 4.3.1.

For each of the above combinations, the networks were evaluated on both the short utterances and the long utterances.

Results. The 4-language network performed at an accuracy of 69.7% on the short utterances and at 82.2% on the long utterances. In comparison, the corresponding 4-language classifier trained on high-quality speech performed at an accuracy of 89.5% on test utterances that were 17.1 seconds long on the average. The confusion matrices for the 4-language networks are shown in Figures 5.11 and 5.12. Not surprisingly, the classification accuracies of all four languages improve substantially on the longer utterances, with Tamil obtaining a perfect score.

label	EN	JA	MA	TA	
EN	65	15	24	11	56.5%
JA	20	79	5	8	70.5%
MA	16	2	77	14	70.6%
TA	6	12	3	92	81.4%

Figure 5.11: Confusion Matrix of 4-language Classifier (Dev. Test): Short Utt.

label	EN	JA	MA	TA	
EN	12	2	4	0	66.7%
JA	1	16	0	0	94.1%
MA	2	0	15	2	78.9%
TA	0	0	0	19	100.0%

Figure 5.12: Confusion Matrix of 4-language Classifier (Dev. Test): Long Utt.

The results of the remaining three experiments are shown in Tables 5.15, 5.16, and 5.17. *#Hid* refers to the number of hidden units that yielded the best development set performance. *Id* refers to the identification performance on the short and long utterances. Columns labeled $EN-L'$, $L-Other$ and $EN-L'-Other$ display the individual language scores.

- It can be seen that the $EN-L'$ classification (Table 5.15) is the least difficult, with performances ranging from 69.0% (English-Farsi) to 87.7% (English-Tamil) for the

short utterances, and from 81.1% (English-German) to 97.3% (English-Tamil) for the "stories". The median accuracies were 79.1% and 86.8% respectively.

- Classification of individual languages against all others ($L-Other$) produces about the same level of performance (Table 5.16), from 63.7% (English-Other) to 86.2% (Mandarin-Other) for the short utterances, and 80.6% (English-Other) to 97.3% (Tamil-Other) for the "stories", with median accuracies of 77.0% and 87.5% respectively.

- English$-L'-Other$ classification is more difficult (Table 5.17), with performances ranging from 53.3% (English-Farsi-Other) to 64.9% (English-Mandarin-Other) for the short utterances, and from 69.2% (English-Vietnamese-Other) to 81.1% (English-Tamil-Other) for the "stories," with median accuracies of 59.1% and 70.6% respectively.

Table 5.15: Results of the English$-L'$ Experiment

Network	*Short Utterances*			*Long Utterances*		
	#Hid	*EN–L' (%)*	*Id (%)*	*#Hid*	*EN–L' (%)*	*Id (%)*
EN-FA	10	70.4–67.6	69.0	7	72.2–94.7	83.8
EN-FR	9	73.9–84.3	79.1	8	72.2–100.0	86.8
EN-GE	7	73.9–72.9	73.4	6	77.8–84.2	81.1
EN-JA	6	74.8–86.6	80.6	10	88.9–94.1	91.4
EN-KO	5	67.8–81.7	74.6	5	88.9–83.3	86.1
EN-MA	10	67.8–82.6	75.0	12	88.9–78.9	83.8
EN-SP	8	74.8–85.6	80.1	10	83.3–100.0	91.9
EN-TA	8	86.1–89.4	87.7	10	94.4–100.0	97.3
EN-VI	7	75.7–85.0	80.2	6	88.9–94.4	91.7

Comparison with Baseline Results. For the short utterances, it is interesting to compare the identification performance of the $EN - L'$ pairs using the combined knowledge-based feature set with that of the baseline label and duration features described in Section 5.4.2. Table 5.18 displays the performance figures for the two feature

Table 5.16: Results of the $L - Other$ Experiment

Network	*Short Utterances*			*Long Utterances*		
	#Hid	*L–O (%)*	*Id (%)*	*#Hid*	*L–O (%)*	*Id (%)*
EN-Other	6	59.1–68.5	63.7	6	61.1–100.0	80.6
FA-Other	10	72.1–70.4	71.2	9	89.5–83.3	86.5
FR-Other	6	75.7–76.9	76.2	10	80.0–88.9	84.2
GE-Other	9	84.7–65.7	75.7	9	89.5–77.8	83.8
JA-Other	7	79.5–75.9	77.7	6	88.2–88.9	88.6
KO-Other	8	67.0–78.7	72.8	6	77.8–88.9	83.3
MA-Other	10	87.2–85.2	86.2	6	89.5–100.0	94.6
SP-Other	6	76.6–80.6	78.5	5	89.5–100.0	94.6
TA-Other	10	85.8–84.3	85.1	6	94.7–100.0	97.3
VI-Other	10	74.8–88.9	81.9	10	88.9–94.4	91.7

Table 5.17: Results of the English–L' – $Other$ Experiment

Network	*Short Utterances*			*Long Utterances*		
	#Hid	*EN– L' –O (%)*	*Id (%)*	*#Hid*	*EN– L' –O (%)*	*Id (%)*
EN-FA-O	7	47.0–53.2–59.8	53.3	10	50.0– 94.7 –62.5	69.8
EN-FR-O	9	45.2–75.7–58.9	59.9	10	66.7– 90.0 –62.5	74.1
EN-GE-O	5	53.9–61.0–63.4	59.4	7	50.0– 89.5 –81.5	73.6
EN-JA-O	7	53.9–68.8–52.7	58.4	7	72.2– 82.4 –56.2	70.6
EN-KO-O	6	54.8–54.1–56.2	55.1	6	55.6– 77.8 –87.5	73.1
EN-MA-O	7	53.0–76.1–66.1	64.9	10	55.6– 78.9 –75.0	69.8
EN-SP-O	8	50.4–71.2–54.5	58.6	8	66.7– 94.7 –43.8	69.8
EN-TA-O	10	60.9–87.6–44.6	64.4	8	55.6–100.0–87.5	81.1
EN-VI-O	5	59.1–72.0–51.8	60.8	6	50.0– 83.3 –75.0	69.2

sets. The third column indicates the percentage change (positive or negative) in the error-rate obtained by using knowledge-based features. Of all the $EN - L'$ pairs, the error-rate on EN-JA actually *increases* by 15.5%, while that on EN-FA increases by a negligible amount. For the remaining seven pairs, using knowledge-based features does improve identification performance, with error-rate decreases ranging from 3.8% (for EN-MA) to 37.3% (for EN-GE).

It is interesting that on the EN-JA task, a single segment-pair feature such as ratio of occurrence of VOC-POVS provided an accuracy of 74.5%, and the addition of 332 other features succeeded in improving the performance to only 80.6%. Similarly, for the EN-TA task, the frequency of occurrence of FRIC alone provides an accuracy of 86%, compared to the 87.7% obtained with all 333 features.

Table 5.18: Baseline and the Combined Feature Sets: A Comparison

Language-Pair	*Baseline (%)*	*Combined Features (%)*	*Error-rate Change (%)*
EN-FA	69.3	69.0	+1.0
EN-FR	72.0	79.1	-25.4
EN-GE	57.6	73.4	-37.3
EN-JA	83.2	80.6	+15.5
EN-KO	65.3	74.3	-26.0
EN-MA	74.0	75.0	-3.8
EN-SP	75.2	80.1	-19.8
EN-TA	85.3	87.7	-16.3
EN-VI	75.8	80.2	-18.2

5.4.11 Final Test

Thus far, all the feature evaluations were performed on the development test set. The final set of 333 features was now used to report results on the final test set. Unlike the development phase, where typically several network configurations (with different numbers of hidden neurons) were trained and evaluated to determine the optimal configuration, a *single* network configuration was trained and evaluated for each experiment in the final test. This network configuration was the one that yielded the best results

on the development set (shown in Tables 5.15, 5.16 and 5.17). Given that the optimal number of hidden neurons, and therefore classification performance is determined by idiosyncrasies of individual test sets, it is reasonable to state that the results obtained on the final test set might not be the optimal result on this data set.

5.4.11.1 Data Sets

Short Utterances. The 500-speaker training set and the 200-speaker development set were combined into a larger final training set. This set had 3834 utterances produced by 700 speakers from 10 languages. The average duration of the utterances was 13.5 seconds. The final test set had 1077 utterances from a new set of 200 speakers. The average duration of the utterances was 13.3 seconds. The utterances in both data sets actually ranged in duration from 1 second to 50 seconds, but these two data sets were collectively referred to as the "short" utterances since the average duration was indeed small when compared to a data set consisting only of "stories".

Long Utterances. The training set for this pair of data sets was identical to the one above, but the test set consisted of only the "stories" in the final test set. There were a total of 178 "stories", one per speaker, ranging in duration from 1 second through 50 seconds, with an average duration of 45.1 seconds.

5.4.11.2 Results

Since training a neural network with additional data almost always improves classification performance, two tests were run on the 10-language task using the final test set but different amounts of training data:

- Test#1: Training on the original training set of 2714 utterances from 500 speakers, and
- Test#2: Training on the combined training set of 3834 utterances from 700 speakers.

Test#1 provided a comparison between performances on the development set and the final test set (using a common training set). It provided insights into the generalization capability of the 333 features. Test#2 served to show the effects of adding more training data on the classification performance (using the final test set as a common test set). Both tests used a 333-60-10 network, the configuration that yielded the best performance on the development set (Section 5.4.10).

The classification performances for Test#1 and Test#2 were 45.0% and 47.3% respectively indicating that addition of more training data does indeed help performance. The drop in performance from 48.5% to 45.0% using a common training set reflects a 6.8% increase in the error-rate when tested on a completely new test set.

The confusion matrices for Test#1 and Test#2 are shown in Figures 5.13 and 5.14 respectively. Note that the addition of training data actually degrades the performance of Mandarin and Spanish (58.7% to 53.2% and 54.1% to 51.4% respectively) while performance on German remains unchanged at 46.5%. The remaining languages show an improvement in performance.

label	EN	JA	MA	TA	SP	KO	VI	FR	GE	FA	
EN	39	14	4	3	8	2	2	13	23	7	33.9%
JA	6	52	10	13	7	0	1	5	11	8	46.0%
MA	8	5	64	2	4	9	7	3	5	2	58.7%
TA	3	7	5	60	5	8	7	4	1	1	59.4%
SP	2	16	3	6	59	4	5	4	6	4	54.1%
KO	7	1	7	6	3	34	14	14	3	6	35.8%
VI	2	5	8	10	8	5	50	6	1	6	49.5%
FR	12	6	10	5	8	12	5	39	2	10	35.8%
GE	13	6	6	1	6	7	8	6	53	8	46.5%
FA	15	4	9	3	12	8	5	8	12	35	31.5%

Figure 5.13: Confusion Matrix for Final Test #1: Original Training Set

Based on the results of Test#1 and Test#2, it was decided to use the larger training set (original training plus development) for all the other tasks. For the 4-language task, the performance accuracy was 66.2% for the short utterances and 85.7% for the long

label	EN	JA	MA	TA	SP	KO	VI	FR	GE	FA	
EN	40	8	5	5	6	2	2	13	18	16	34.8%
JA	3	61	8	10	8	0	2	5	9	7	54.0%
MA	5	4	58	5	4	6	7	7	5	8	53.2%
TA	1	7	3	63	7	7	8	3	1	1	62.4%
SP	3	19	1	6	56	2	8	3	4	7	51.4%
KO	6	4	5	8	3	39	8	14	3	5	41.1%
VI	1	4	8	6	9	3	55	6	1	8	54.5%
FR	7	11	13	1	8	8	7	43	2	9	39.4%
GE	10	7	8	1	9	5	5	6	53	10	46.5%
FA	6	5	7	5	13	6	7	6	15	41	36.9%

Figure 5.14: Confusion Matrix for Final Test #2: Augmented Training Set

utterances. Confusion matrices are given in Figures 5.15 and 5.16 respectively.

label	EN	JA	MA	TA	
EN	72	27	10	6	62.6%
JA	19	73	12	9	64.6%
MA	24	6	72	7	66.1%
TA	4	17	7	73	72.3%

Figure 5.15: Confusion Matrix for 4-language Classifier (Final Test): Short Utt.

label	EN	JA	MA	TA	
EN	16	3	0	0	84.2%
JA	2	17	0	0	89.5%
MA	2	1	15	0	83.3%
TA	0	2	0	12	85.7%

Figure 5.16: Confusion Matrix for 4-language Classifier (Final Test): Long Utt.

The results on the $EN - L'$, $L - Other$ and $EN - L' - Other$ tasks are summarized in Figures 5.17, 5.18 and 5.19 respectively. Figure 5.20 summarizes the results of all

the experiments on the final test set. The bars for the English–L', L – *Other* and English–L' – *Other* tasks represent median scores.

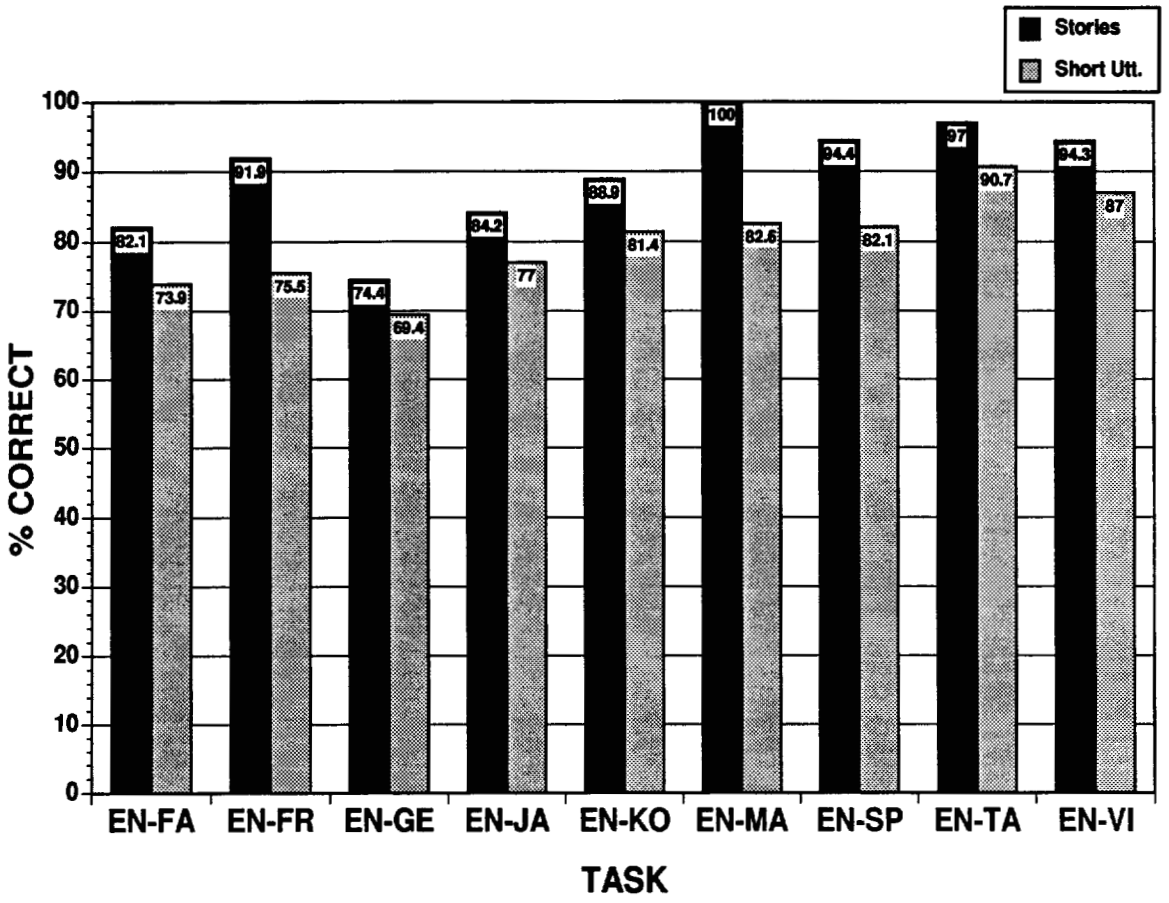

Figure 5.17: Results of the English–L' Final Test

5.5 Summary

This chapter described research into speaker-independent automatic language identification using a ten-language telephone speech corpus.

A neural network-based broad phonetic segmentation algorithm for telephone speech was developed using just PLP-based spectral features. Features based on linguistic

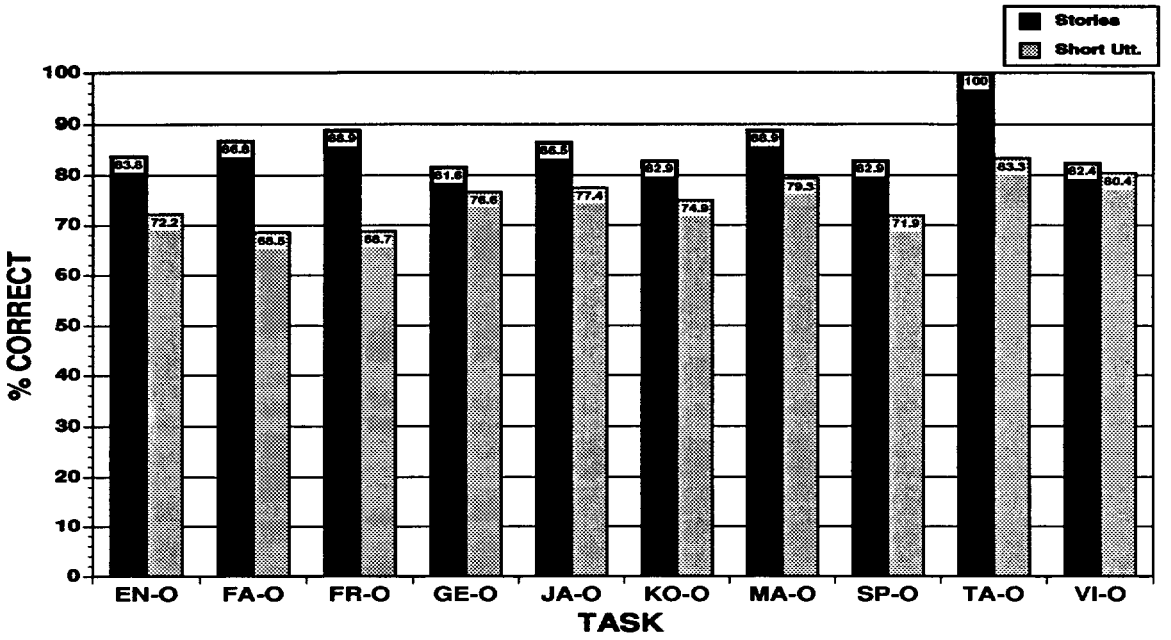

Figure 5.18: Results of the $L - Other$ Final Test

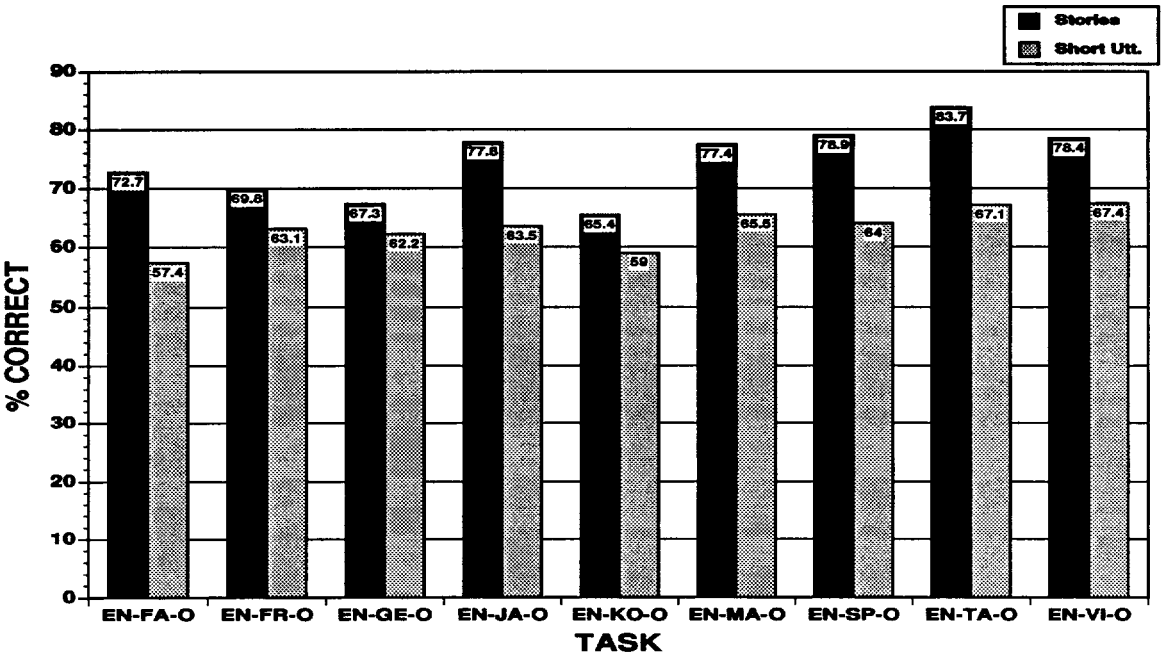

Figure 5.19: Results of the English–L' – $Other$ Final Test

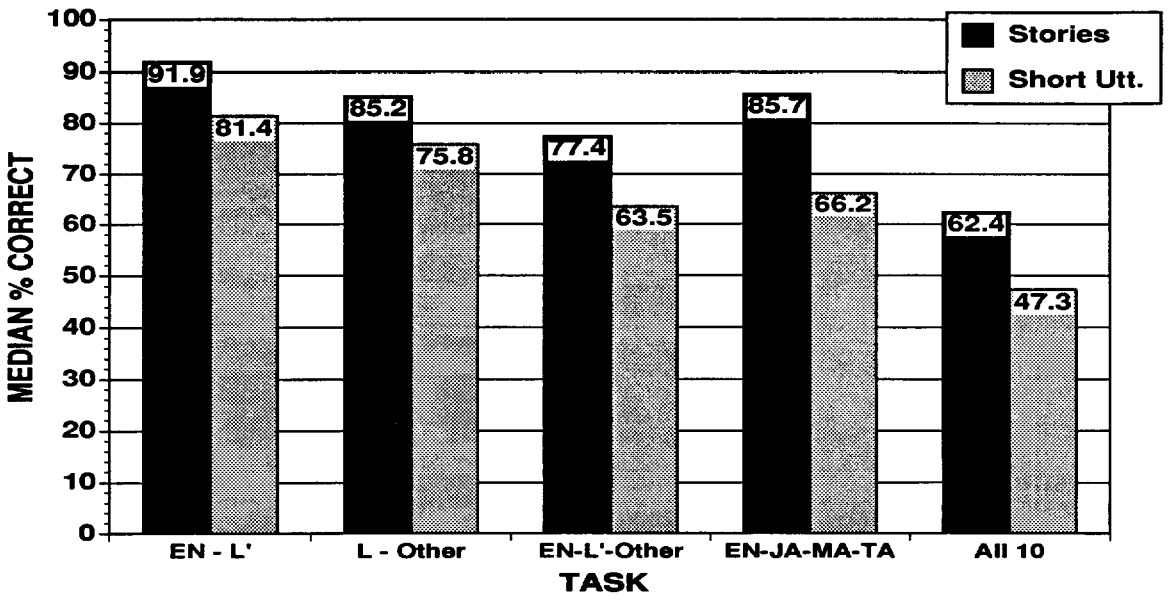

Figure 5.20: Summary of Language Identification Results on the Final Test Set

knowledge and on exploratory statistical analyses of the languages were measured on the broad-phonetic segment sequences, and input to a second neural network which performed the language classification. The types of features examined ranged from simple segment duration and label information (the baseline feature set) to segment-pair frequencies, segment-triple frequencies, duration statistics of segments, and intra- and inter-segment pitch variations. In addition, PLP-based spectral features, with and without segmentation, were also examined. On the ten-language task, it was found that features based on segment-pairs performed the best and those based on pitch variation, the worst. In order to remove redundancies among the features, the feature sets were reduced using principal component analyses and visual inspection of the boxplots. In most cases, the language identification performance using just the principal components was comparable to that with the full-blown feature set, with substantial reductions in the dimensions of the input feature space. Language identification performance on a combined feature set was then contrasted with that on a combination of principal components of each feature

set. The full-blown feature set performed slightly better. Using this feature set, language classification experiments with smaller combinations of languages were conducted. Not surprisingly, language identification performance on constrained tasks (fewer languages) was reasonably high, and dropped steeply as more languages were added. Evaluation on the longer "story" utterances sets provided significant performance improvements for almost all the languages.

On the two-language task, single features such as the ratio of occurrence of the segment-pair VOC-POVS and the frequency of occurrence of FRIC, proved to be most useful for the EN-JA and EN-TA tasks, respectively. The language identification performance with each of these features was only slightly worse than that obtained with all 333 features on their respective tasks. Spectral features were examined for the EN-JA task alone. Frame-based PLP features with no segmentation proved to be inferior to segment-based features in language identification performance on this task. On the $EN - L'$ task, comparison of the baseline results with those of the knowledge-based features showed that EN-JA was the lone pair that did significantly better on the baseline features than on the 333 combined features!

Since the development set was extensively used to optimize the feature set, results were reported on a new hitherto untouched test set, the final test set. For the final test, the original training set was augmented with the development set to provide more training data to the classifier. With the exception of Mandarin, Spanish and German, the individual language performances on the final test set were better than the corresponding development set results.

Chapter 6

Conclusion

6.1 Comparisons with Recent Work

To place the results described in Chapter 5 in proper perspective, it is pertinent to compare them to those of two recent studies that have also used the OGI Multi-language Telephone Speech Corpus (with the same training, development test and final test divisions).

Marc Zissman at MIT Lincoln Laboratory has reported results of a HMM-based approach to automatic language identification [Zis93]. Using continuous observation, ergodic hidden Markov models (HMMs) with tied Gaussian observation probability densities, he obtained an identification accuracy of 46.0% on the development test set for the ten-language task. He found that performance of a Gaussian mixture classifier (a single-state HMM) was comparable to that of multi-state HMMs, indicating that the sequential modeling capabilities of HMMs were not exploited. His results on the English$-L'$ and $L - Other$ tasks were also comparable to our results. He did not examine the English$-L' - Other$ task.

Timothy Hazen and Victor Zue at MIT have reported results on a segment-based approach to automatic language identification, designed around a formal probabilistic framework [HZ93]. Using probablistic models for the phonotactic, prosodic and acoustic properties of the different languages in the corpus, they obtained an identification accuracy of 47.7% on the development test set for the ten-language task. Results on the English$-L'$, $L - Other$ and English$-L' - Other$ tasks were not reported.

Results of both these studies are in close agreement with the 48.5% reported in Chapter 5 using knowledge-based features and artificial neural networks. While Zissman used a completely different approach (HMMs and tied Gaussian probability densities), Hazen and Zue's approach, like the research described in this dissertation, was inspired by the work of House and Neuberg [HN77], who proposed that languages can be differentiated based on sequential constraints on broad phonetic categories. The fact that two other approaches to the same problem, using the same corpus, have arrived at comparable results, is an indication of the inherent difficulty of the problem. It is also significant that the HMM approach, which did not utilize any (broad or fine) phonetic transcriptions of speech, produced results comparable to the other two approaches that relied on broad phonetic transcriptions.

6.2 Contributions

The two main contributions of this dissertation are the development of a ten-language telephone speech corpus and the detailed examination of an approach to automatic language identification that relies only on the properties of seven broad phonetic categories.

6.2.1 OGI Multi-language Telephone Speech Corpus (OGI_TS)

Before the development of the OGI Multi-language Telephone Speech Corpus (OGI_TS), the field of language identification had a relatively small number of published papers reflecting twenty years of research, performed mainly by government agencies and companies on contract to government agencies. Much of the research was classified, and it was difficult to make objective comparisons between the results of the studies that were published, since they used different corpuses, and the languages used and other experimental details were not publicly available.

Table 6.1: Participants of the First NIST LangId Evaluation

Name	*Location*
Emerson & Stein	San Diego, CA
GTE	Baltimore, MD
ITT	San Diego, CA
Lockheed-Sanders	Nashua, NH
MIT Lincoln Laboratory	Cambridge, MA
MIT Laboratory for Computer Science	Cambridge, MA
Oregon Graduate Institute (OGI)	Portland, OR
Renssaeler Polytechnic Institute (RPI)	Troy, NY

The free availability[1] of the OGI Multi-language Telephone Speech Corpus, developed specifically to support the research described in this dissertation, has substantially altered this landscape. For example, a new initiative in automatic language identification is now under way in the United States. In March 1993, NIST designated OGI_TS as the standard for evaluating language identification algorithms. Researchers at eight different sites across the United States (see Table 6.1) are working on different approaches to automatic language identification using OGI_TS. The evaluation of the different approaches is being conducted by NIST. The tasks consist of identification of all ten languages and the language groups $L-Other$, English$-L'$, and English$-L'-Other$ described in Chapter 5. At the time of writing, results of the first NIST language identification evaluation held in June 1993 were yet to be formally published.

It is clear that the interaction and exchange of ideas resulting from different sites working on different approaches to the problem using the same speech corpus will result in significant advances in the area of automatic language identification, and the development of techniques to assess and compare different language identification methods.

[1]The corpus is distributed free of charge only to universities, not-for-profit organizations and research laboratories funded by the U.S. government.

6.2.2 Detailed Examination of the Broad Phonetic Approach

While a majority of the approaches to automatic language identification over the past two decades have been frame-based statistical approaches [CI82, Ive86, Foi86, GMW89, Zis93], very few researchers [LE80] have followed up on House and Neuberg's seminal work [HN77] in which they proposed that languages could be distinguished solely on the basis of sequential constraints on broad phonetic categories. The broad phonetic approach described in this dissertation and Hazen and Zue's work described above are two recent studies that have expanded on House and Neuberg's work.

One of the objectives of the dissertation was to determine the extent to which features derived from sequences of broad phonetic categories would help in distinguishing between ten languages using telephone speech. As shown in Chapter 5, features based on individual broad phonetic categories, as well as pairs and triples and ratios combined perform at an accuracy of 47.3% on the utterances of the final test set (average duration: 13.4 seconds). The results for the English$-L'$, $L-Other$ and English$-L'-Other$ tasks range from 69.4% to 90.7%, 68.5% to 83.3%, and 57.4% to 67.4% respectively, on the short and long utterances. These results are much better than chance performance (10% for the ten language task, 50% for the two-language task and 33.3% for the three-language task) and indicate that sequences of broad phonetic categories do reflect some degree of language discriminatory ability. Comparison of the baseline and knowledge-based feature results on the pairwise language identification task indicate that knowledge-based features do perform better than features with minimal linguistic knowledge encoded in them. Further, the discriminative power of the knowledge-based features does increase with duration of the utterances, as shown by the results for the long utterances (Section 5.4.11.2). As expected, there was a degradation in performance in moving from high-quality speech to telephone speech for the four-language task.

These results are encouraging and validate House and Neuberg's hypothesis that sequential constraints on broad phonetic categories have the ability to distinguish between languages. However, the level of identification performance achieved with broad

phonetic features still leaves considerable room for improvement. Although the research described in this dissertation may not have exhaustively searched the broad phonetic category feature space, it has explored enough of it to determine that broad phonetic category information alone is not sufficient for the development of viable language identification systems. By "viable" I mean systems that could be useful on a day-to-day basis. The level of accuracy that such systems demand requires approaches that mimic the way human beings possibly perform language identification: by cuing onto specific phonemes or phoneme sequences and words in the language.

6.3 Future Work

Within the broad phonetic framework, there is room for improvement in two areas:

- increased accuracy of the broad phonetic category segmentation algorithm. This can be achieved by increasing the training data (involves more hand-labeling) and by training a boundary classifier to reduce the number of boundary errors. Given that all of the language classifier features are based on sequences of broad phonetic categories, it is likely that improvements in the segmenter performance would translate to significant improvements in language identification performance.

- improved prosodic features. The pitch variation features examined in this dissertation performed the poorest of all feature sets. Features incorporating intonation contours and tone variations are likely to help in distinguishing between the tone languages and the non-tone languages in the corpus.

Assuming that exploiting differences between languages at the phonemic and phonetic levels is the key to more accurate language identification, the availability of a phonemically or phonetically labeled corpus of data is essential. There is current work at the Oregon Graduate Institute in two areas: providing a phonemic transcription of the utterances in the corpus and exploring phonemic approaches to automatic language identification.

6.3.1 Phonemic Transcriptions

Providing an accurate phonetic transcription of speech is a time-consuming task and brings up the question of precision versus accuracy. The more precise the phonetic transcription, the greater the likelihood of transcription errors. It also requires the expertise of linguists and phoneticians who are well-versed in the phonology and the acoustic-phonetic structure of the languages involved and the use of a phonetic notation convention such as the International Phonetic Alphabet (IPA). In view of these difficulties involved in arriving at an accurate *phonetic* transcription of speech, the OGI Speech Laboratory is currently producing *phonemic* transcriptions of the languages in OGI_TS. A phoneme is the smallest distinguishable unit of sound in a language. A phone is the articulatory realization of a phoneme. A phonemic transcription is less detailed than a phonetic transcription in the sense that different articulatory realizations of the same phoneme (called allophones) are essentially ignored. This approach has the following advantages:

- it minimizes problems of precision vs. accuracy since robust phonemic transcriptions can generally be obtained with reasonable training, and
- the transcribers need not be expert linguists and phoneticians; fluent native speakers with some knowledge of the phonology of the language can be trained for the transcription task.

However, one of the problems with this approach is that the same symbol is used to represent different sounds across the languages. This might make it difficult to train a classifier since the symbols do not represent consistent mappings to sounds across languages.

Two native speakers of each language, familiar with its phonology, are being trained to provide phonemic transcriptions to the utterances in OGI_TS. For each language, each transcriber labels slightly more than half of the utterances in that language. This ensures that some of the utterances get transcribed by two different transcribers. These duplicate

transcriptions are used to determine inter-transcriber reliability for that language. Care is being taken to have a uniform set of labeling conventions across the different languages to avoid ambiguities and phoneme label conflicts.

This transcription effort is clearly an arduous and time-consuming process, but the benefits of a phonemically labeled corpus of multi-lingual speech fully justify the large investment of time and effort. Such a corpus would be invaluable in pursuing a phonemic approach to automatic language identification and would serve as a common database for assessment and comparison of techniques.

6.3.2 Phoneme-based Approaches

Preliminary results are in from a phoneme-based approach to automatic language identification for the English-Japanese task [Ber93, MBA+93]. The system consists of two stages:

- PLP features derived from the signal are used to perform a frame-based phoneme classification for each language.
- Unigram and bigram features derived from the frame-by-frame classifier outputs are fed into a language classifier network.

Figure 6.1 displays a schematic block diagram of the system. The English phonemic front-end uses 39 phonemes, while the Japanese front-end has 25 phonemes. The English phonemic front-end currently performs at an accuracy of 48%, and the Japanese front-end at an accuracy of 46%. The unigram features consist of the average output activation, the maximum output activation, and the variation in the output activation for each of the 64 ($= 39 + 25$) phonemes. The bigram features consist of those transition probabilities between pairs of phonemes whose average values over the Japanese and English training sets differed the most. While the English-Japanese classification performances using the unigram and bigram features separately (82.3% and 79.3% respectively) were inferior to that of the baseline sliding window approach described in section 5.4.2.3 (83.2%),

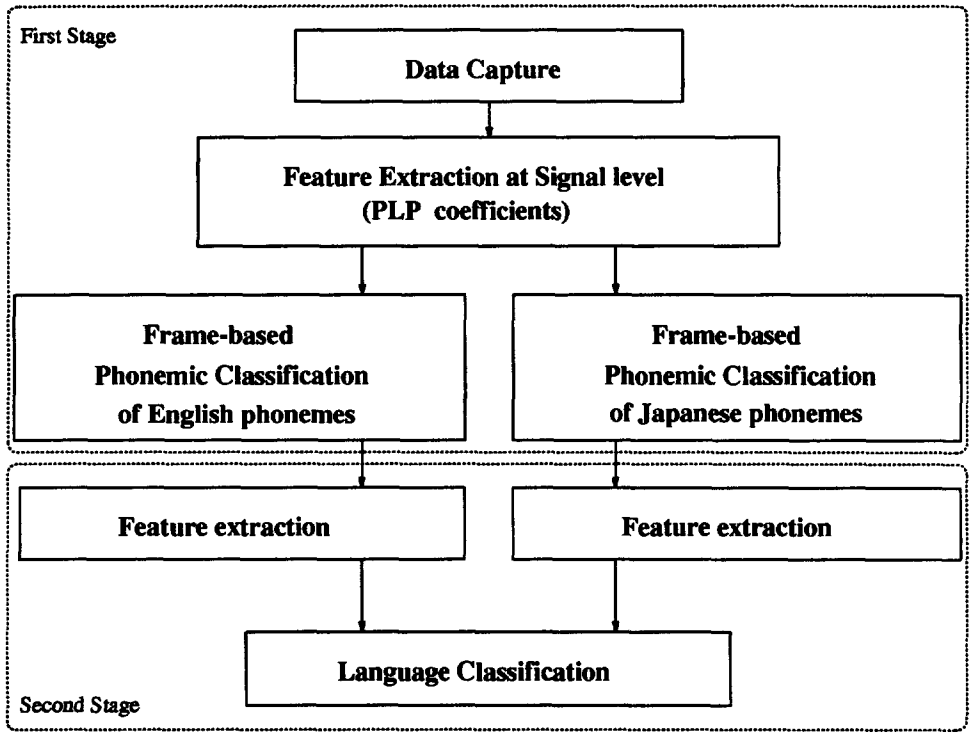

Figure 6.1: Modules of the Phoneme-based Language Identification System

classification performance improved to 86.3% when the unigram and bigram features were combined in a single network.

This is an encouraging result, given the relatively poor performances of the English and Japanese phonemic front-ends. Apparently, the low accuracy of the phonemic classifiers is offset by the more detailed information that they provide. Work is under way to extend this approach to the full ten-language set and to improve the classification accuracy of the phonemic front-ends by the addition of more hand-labeled data.

Bibliography

[Ass89] International Phonetic Association. Report on the 1989 Kiel Convention. *Journal of the International Phonetic Association*, 19(2):67–80, 1989.

[Atk68] K. Atkinson. Language identification from nonsegmental cues. *Journal of the Acoustical Society of America*, 44:378(A), 1968.

[BC89] E. Barnard and R. A. Cole. A neural-net training program based on conjugate-gradient optimization. Technical Report CSE 89-014, Department of Computer Science, Oregon Graduate Institute of Science and Technology, 1989.

[BCVA91] E. Barnard, R. A. Cole, M. P. Vea, and F. A. Alleva. Pitch detection with a neural-net classifier. *IEEE Transactions on Signal Processing*, 39(2):298–307, February 1991.

[Ber93] K. M. Berkling. Automatic language identification using statistical features derived from phoneme classifiers. Presented at the 1993 OGI CSE Student Research Symposium, May 1993.

[BHH78] G. E. P. Box, W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters.* John Wiley, New York, 1978.

[Bur92] D. Burnett. Toward multi-language pitch tracking for telephone speech. Presented at the 1992 OGI CSE Student Research Symposium, May 1992.

[C+92] R. A. Cole et al. Workshop on spoken language understanding. Technical Report CS/E 92-014, Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, 1992.

[CFC91] J. W. Creekmore, M. A. Fanty, and R. A. Cole. A comparative study of five spectral representations for speaker-independent phonetic recognition. In *Proceedings of the 25th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, November 1991.

[CI82] D. Cimarusti and R. B. Ives. Development of an automatic identification system of spoken languages: Phase 1. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 82*, Paris, France, May 1982.

[CL92] B. Chigier and H. C. Leung. The effects of signal representations, phonetic classification techniques and the telephone network. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.

[Com90] B. Comrie, editor. *The World's Major Languages.* Oxford University Press, New York, 1990.

[Cry87] D. Crystal. *The Cambridge Encyclopedia of Language*, pages 280–339. Cambridge University Press, New York, 1987.

[Ead82] S. J. Eady. Differences in the F_0 patterns of speech: Tone language versus stress language. *Language and Speech*, 25(1):29–42, 1982.

[FC91] M. A. Fanty and R. A. Cole. A comparison of DFT, PLP and Cochleagram for alphabet recognition. In *Proceedings of the 25th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, November 1991.

[FCR92] M. A. Fanty, R. A. Cole, and K. Roginski. English alphabet recognition with telephone speech. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, San Mateo, CA, 1992. Morgan Kaufmann Publishers.

[FDGM86] W. Fisher, G. R. Doddington, and K. Goudie-Marshall. The DARPA speech recognition research database: Specification and status. In *Proceedings DARPA Speech Recognition Workshop*, pages 93–100, February 1986.

[Foi86] J. T. Foil. Language identification using noisy speech. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 86*, Tokyo, Japan, 1986.

[FPC92] M. A. Fanty, J. Pochmara, and R. A. Cole. An interactive environment for speech recognition research. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.

[GJD87] C. Grover, D. G. Jamieson, and M. B. Dobrovolsky. Intonation in English, French and German: Perception and production. *Language and Speech*, 30(3):277–295, 1987.

[GMW89] F.J. Goodman, A.F. Martin, and R.E. Wohlford. Improved automatic language identification in noisy speech. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 89*, Glasgow, Scotland, May 1989.

[Gop90] M. Gopalakrishnan. Segmenting speech into broad phonetic categories using neural networks. Master's thesis, Oregon Graduate Institute of Science & Technology, Beaverton, OR, August 1990.

[Her90] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87:1738–1752, April 1990.

[HN77] A. S. House and E. P. Neuberg. Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *Journal of the Acoustical Society of America*, 62(3):708–713, 1977.

[HZ93] T. J. Hazen and V. W. Zue. Automatic language identification using a segment-based approach. In *Proceedings 3rd European Conference on Speech Communication and Technology (Eurospeech 93)*, Berlin, Germany, September 1993.

[IIK90] H. Irii, K. Itoh, and N. Kitawaki. Multilingual speech database for evaluating quality of digitized speech. In *Proceedings International Conference on Spoken Language Processing 90*, pages 1025–1028, Kobe, Japan, 1990.

[Ive86] R. B. Ives. A minimal rule AI expert system for real-time classification of natural spoken languages. In *Proceedings 2nd Annual Artificial Intelligence and Advanced Computer Technology Conference*, Long Beach, CA, April-May 1986.

[JFC91] R. D. T. Janssen, M. A. Fanty, and R. A. Cole. Speaker-independent phonetic classification of continuous english letters. In *Proceedings 1991 International Joint Conference on Neural Networks*, Seattle, WA, July 1991.

[JKBS90] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. In

Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 90, Albuquerque, NM, April 1990.

[Jun89] J. C. Junqua. *Towards Robustness in Isolated-word Automatic Speech Recognition.* PhD thesis, University of Nancy I, 1989.

[KKWE92] S. C. Kwasny, B. L. Kalman, W. Wu, and A. M. Engebretson. Identifying language from speech: An example of high-level, statistically-based feature extraction. In *Proceedings 14th Annual Conference of the Cognitive Science Society*, 1992.

[LD74] R. G. Leonard and G. R. Doddington. Automatic language identification. Technical Report RADC-TR-74-200, Air Force Rome Air Development Center, August 1974.

[LD75] R. G. Leonard and G. R. Doddington. Automatic language identification. Technical Report RADC-TR-75-264, Air Force Rome Air Development Center, October 1975.

[LD78] R. G. Leonard and G. R. Doddington. Automatic language discrimination. Technical Report RADC-TR-78-5, Air Force Rome Air Development Center, January 1978.

[LE80] K.P. Li and T. J. Edwards. Statistical models for automatic language identification. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 80*, Denver, CO, April 1980.

[Leo80] R. G. Leonard. Language recognition test and evaluation. Technical Report RADC-TR-80-83, Air Force Rome Air Development Center, March 1980.

[LKS86] L. Lamel, R. Kassel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proceedings DARPA Speech Recognition Workshop*, pages 100–110, February 1986.

[LRH90] T. K. Leen, M. Rudnick, and D. Hammerstrom. Hebbian feature discovery improves classifier efficieny. In *Proceedings of the International Joint Conference on Neural Networks*, San Diego, CA, June 1990.

[Lyo82] R. F. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 82*, May 1982.

[Mad84] I. Maddieson. *Patterns of Sounds.* Cambridge University Press, Cambridge, Great Britain, 1984.

[Mak75] J. Makhoul. Spectral Linear Prediction: Properties and Applications. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23:283–296, 1975.

[MBA+93] Y. K. Muthusamy, K. M. Berkling, T. Arai, R. A. Cole, and E. Barnard. A comparison of approaches to automatic language identification using telephone speech. In *Proceedings 3rd European Conference on Speech Communication and Technology (Eurospeech 93)*, Berlin, Germany, September 1993.

[MC92a] Y. K. Muthusamy and R. A. Cole. A segment-based automatic language identification system. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, San Mateo, CA, 1992. Morgan Kaufmann Publishers.

[MC92b] Y. K. Muthusamy and R. A. Cole. Automatic segmentation and identification of ten languages using telephone speech. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.

[MCO92] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.

[NUS92] S. Nakagawa, Y. Ueda, and T. Seino. Speaker-independent, text-independent language identification by HMM. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.

[RF91] T. Robinson and F. Fallside. A recurrent error propagation network speech recognition system. *Computer Speech and Language*, 5:259–274, July 1991.

[Rog91] K. Roginski. A neural network phonetic classifier for telephone speech. Master's thesis, Oregon Graduate Institute of Science & Technology, Beaverton, OR, November 1991.

[SAG91] M. Savic, E. Acosta, and S. K. Gupta. An automatic language identification system. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing 91*, Toronto, Canada, May 1991.

[Sla88] M. Slaney. Lyon's Cochlear Model. Technical report, Apple Computer Inc., 1988.

[Sta91] Statistical Sciences, Inc., Seattle, Washington. *S-PLUS User's Manual Volume 2*, September 1991.

[Sug91a] M. Sugiyama. Automatic language recognition using acoustic features. Technical Report TR-I-0167, ATR Interpreting Telephony Research Laboratories, 1991.

[Sug91b] M. Sugiyama. Automatic language recognition using acoustic features. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing 91*, Toronto, Canada, May 1991.

[Zis93] M. A. Zissman. Automatic language identification using gaussian mixture and hidden markov models. In *Proceedings International Conference on Acoustics, Speech and Signal Processing 93*, Minneapolis, MN, April 1993.

Appendix A

Telephone Speech Corpus Recording Protocol

Given below is a transcript of the recording instructions and prompts that callers to our toll-free phone number heard. For languages other than English, the instructions and prompts were translated into that language by a native speaker. The following conventions are observed in the text of the transcript:

- Sentences and phrases spoken only in English are in italics.
- The recording time is given in parentheses following each prompt or question.
- **LANGUAGE** is a placeholder for one of the ten languages.
- **TONE** refers to the machine-generated tone that was used to signal the start of the final 10 seconds of the 1 minute elicited free-speech response.
- The two-letter code in bold (e.g., **dw**) beside each prompt is the acronym of the response. **sb** refers to portion of the "story" *before* the **TONE** and **sa** is the portion of the "story" *after* the tone.

A.1 Transcript

Thank you for calling the Oregon Graduate Institute language database. We are currently recording speech in American English, Farsi, French, German, Korean, Mandarin Chinese, Japanese, Spanish, Tamil and Vietnamese.

To record in German, please press 1 now.

To record in Japanese, please press 2 now.

To record in Korean, please press 3 now.

To record in Mandarin, please press 4 now.

To record in Spanish, please press 5 now.

To record in Tamil, please press 6 now.

To record in Vietnamese, please press 7 now.

To record in French, please press 8 now.

To record in English, please press 9 now.

To record in Farsi, please press 0 now.

You have selected the Oregon Graduate Institute **LANGUAGE**-language database. We are studying the different languages of the world. To do this, we need to record samples of speech from native speakers of **LANGUAGE**. Please respond to the following questions and instructions in **LANGUAGE** only. Please wait for the beep before speaking.

- What is your native language? (3 seconds) (**nl**)
- What language do you speak most of the time? (3 seconds) (**cl**)
- Please recite the 7 days of the week. (8 seconds) (**dw**)
- Please count from 0 through 10. (11 seconds) (**nm**)

We will now ask you to tell us something about your home-town. If you do not wish to talk about your home-town, you can talk about any city of your choice. Please wait for the beep before speaking.

- Tell us something that you like about your home-town. (10 seconds) (**hl**)
- Tell us about the climate in your home-town. (10 seconds) (**hc**)

We will now ask you to describe certain events and locations. If you do not wish to describe them, feel free to make some up fictional descriptions. Please wait for the beep before speaking.

- Describe the room that you are calling from. (12 seconds) (**rm**)
- Describe what you had for your most recent meal. (10 seconds) (**ml**)

We now want you to talk for a longer period of time. We do not care what you say as long you keep talking. You can tell us anything about yourself, your hobbies and interests, the city that you live in, and the sports that you like. Or you can make up a story, tell a fairy-tale or recite a poem. You will have 1 minute to speak. We will now give you 10 seconds to think about what to say.

10 second pause

Please begin talking at the beep. You will hear a tone like this **TONE** when you have 10 seconds left. Please continue talking and finish your story. (1 minute) (**sb** and **sa**)

Thank you for calling. We appreciate your help.

Appendix B

Corpus Statistics

B.1 Corpus Distribution Sites

Table B.1 lists the research sites that have received either or both the four-language high-quality speech corpus (OGI_HQ) and the ten-language telephone speech corpus (OGI_TS). Sites marked with a ‡ received both the high-quality and the telephone speech corpuses, while the one marked with a † received just the high-quality speech corpus.

B.2 Four-language High-quality Speech Corpus

B.2.1 Speaker and Utterance Information by Language

Tables B.2, B.3, B.4 and B.5 display speaker and utterance information for the English, Japanese, Mandarin and Tamil portions of the corpus, respectively. Not all of the speakers in Japanese and Mandarin were used in the segmentation and classification experiments. The column headings represent:

- *SpkrID* - Speaker initials (of the form AAN or AAAN where 'A' is a letter of the alphabet and 'N' is a digit $0-9$ to disambiguate identical initials between speakers of the same gender). Initials with asterisks next to them represent speakers that were not used in any of the experiments.
- *Sex* - Speaker gender (m or f)
- *RecDate* - Speaker recording date

Table B.1: OGI Multi-lingual Speech Corpus Distribution Sites

Site	*Location*
AT&T	Murray Hill, NJ
Boston Univ.	Boston, MA
Dragon Systems, Inc.‡	Newton, MA
Emerson & Stern	San Diego, CA
IDA	Princeton, NJ
IDIAP	Martigny, Switzerland
LDC	Pittsburg, PA
LIMSI CNRS	Orsay, France
MIT‡	Cambridge, MA
MIT Lincoln Labs‡	Cambridge, MA
MITRE Corporation†	VA
NC State Univ.‡	Raleigh, NC
NSA‡	Fort Meade, MD
Old Dominion Univ.	Norfolk, VA
NIST	Gaithersburg, MD
US West	Boulder, CO
Washington Univ.	St. Louis, MO

- *BirthDate* - Speaker birth date
- *Origin* - state(s) or country where the speaker spent most of his/her youth
- *#RecUtt* - Number of utterances recorded by the speaker
- *#HndLabel* - Number of utterances for which broad phonetic transcriptions were provided

B.2.2 Segmenter Training, Development and Final Test Sets

Tables B.6, B.7 and B.8 display the speakers and utterances from each language chosen for the segmenter training, development and final test sets respectively. The following naming conventions are used for the utterances provided by each speaker.

- s1 through sN - conversational utterances, where 'N' ranged from 7 through 19
- q1 and q2 - questions
- 'dw' - days-of-the-week
- 'nm' - numbers 0 through 10

B.2.3 Classifier Training, Development and Final Test Sets

Tables B.9, B.10 and B.11 display the speakers used in the language classification training, development test and final test sets, respectively. All utterances from each speaker were used.

Table B.2: OGI_HQ: Speaker and Utterance Information for English

SpkrID	*Sex*	*BirthDate*	*RecDate*	*Origin*	*#RecUtt*	*#HndLabel*
bbf0	f	1/24/64	4/2/91	Guam	20	5
cp0	f	1/3/55	4/1/91	OR	20	5
dcb0	m	11/29/68	3/20/91	VA	20	5
dcd0	m	1/13/57	3/29/91	OR	20	5
jbc0	f	6/1/46	4/2/91	VA,CA	20	5
jjs0	m	3/9/57	3/28/91	MO	20	5
jw1	f	6/2/52	4/9/91	OR	20	5
ka0	f	11/3/66	3/26/91	OR	20	5
kdr0	m	1/3/63	3/27/91	MN	20	5
kw0	f	7/17/56	3/26/91	OR	20	0
lvl0	f	10/28/61	3/18/91	NY	20	5
mm1	f	8/10/55	3/27/91	IA	20	5
njh0	m	4/4/64	3/27/91	NY	20	5
pmb0	m	1/9/52	3/28/91	WA	20	5
rms0	m	7/5/48	3/29/91	NY	20	5
sjs0	f	11/7/51	4/10/91	NYC,NJ	20	5
slb0	m	9/9/56	3/25/91	ME	20	5
ss0	f	5/7/64	4/1/91	OR	20	5
tkl0	m	12/24/55	3/25/91	NJ	20	5
vcw0	m	11/11/50	3/28/91	OR	20	0

Table B.3: OGI_HQ: Speaker and Utterance Information for Japanese

SpkrID	*Sex*	*BirthDate*	*RecDate*	*Origin*	*#RecUtt*	*#HndLabel*
cy0	f	6/3/58	11/27/90	Japan	20	5
ea0	f	3/8/56	11/20/90	Japan	20	5
fi0	f	3/4/57	12/28/90	Japan	20	5
ft0	m	7/31/53	12/6/90	Japan	20	5
fy0	m	3/22/55	3/1/90	Japan	23	5
ho0	f	1/24/57	3/12/90	Japan	20	5
hw0	m	1/5/58	3/14/90	Japan	20	5
kk0*	m	2/15/65	1/6/92	Japan	20	0
ks0	f	4/14/31	6/15/90	Japan	10	5
ku0	m	6/5/72	3/20/90	Japan	10	5
mi0	m	1/5/57	12/20/90	Japan	20	5
mk0	f	10/25/63	12/30/91	Japan	20	0
mm0	f	5/4/56	12/12/90	Japan	20	5
mn0	f	2/28/50	6/15/90	Japan	10	5
mn0*	m	10/23/55	3/1/90	Japan	20	5
mt0	f	2/23/60	12/6/90	Japan	20	5
mw0*	f	12/4/67	1/6/92	Japan	20	0
nu0	f	7/26/75	3/20/90	Japan	10	5
st0	f	6/19/58	6/15/90	Japan	10	5
st0	m	12/13/73	7/18/91	Japan	20	5
ta0	m	2/6/56	11/20/90	Japan	20	5
tf0	m	6/7/74	7/18/91	Japan	20	5
tu0*	m	1/31/59	3/1/90	Japan	20	5
yi0	f	12/10/60	2/8/91	Japan	20	5
ys0*	m	2/4/40	2/28/90	Japan	20	5

Table B.4: OGI_HQ: Speaker and Utterance Information for Mandarin Chinese

SpkrID	*Sex*	*BirthDate*	*RecDate*	*Origin*	*#RecUtt*	*#HndLabel*
cd0	m	3/26/66	11/9/90	China	20	5
dl0*	m	11/11/55	11/9/90	China	20	5
fl0	f	1/4/40	11/18/90	China	20	5
gg0	f	7/28/62	11/12/90	China	20	5
gl0	m	4/15/65	12/10/90	China	20	5
jh0	f	3/13/62	11/20/90	Taiwan	20	5
jq0	m	8/1/64	5/10/90	China	10	5
jw0	f	9/23/61	11/19/90	China	20	5
ll0	f	7/23/73	5/6/90	China	20	5
ls0	m	10/13/63	11/10/90	China	20	5
lz0	f	4/8/66	11/18/90	China	20	5
lz0	m	2/16/65	4/28/90	China	20	5
ny0	m	4/2/58	11/8/90	China	20	5
ql0*	m	10/17/37	11/18/90	China	20	0
qx0	f	12/29/66	11/16/90	China	20	5
sg0	f	12/25/44	11/27/90	China	20	0
ws0	m	11/29/58	5/11/90	China	20	5
xh0	f	2/11/65	11/2/90	China	20	5
xw0	m	10/17/62	11/16/90	China	20	0
yl0	m	3/1/40	6/13/90	china	20	0
yp0	f	8/1/69	11/10/90	China	20	5
yt0	m	4/1/53	5/12/90	China	20	5
zb0	m	2/8/66	11/9/90	China	20	0
zl0	m	3/30/68	11/12/90	China	20	0

Table B.5: OGI_HQ: Speaker and Utterance Information for Tamil

SpkrID	*Sex*	*BirthDate*	*RecDate*	*Origin*	*#RecUtt*	*#HndLabel*
akg0	m	6/4/67	11/23/90	India	20	6
dr0	m	8/12/63	11/30/90	India	24	5
hs0	f	3/27/68	7/24/90	India	20	5
js0	f	5/25/65	12/31/90	India	20	5
jv0	f	5/6/52	12/5/90	India	20	5
la0	f	11/21/51	1/3/91	India	20	5
mg0	m	11/13/63	2/28/90	India	10	5
ng0	f	9/12/62	11/26/90	India	20	5
nm0	m	12/22/19	12/1/90	India	20	5
rps0	m	5/18/60	11/30/90	India	24	5
rr0	m	6/14/63	3/6/90	India	20	5
rs0	m	11/16/59	1/31/91	India	20	0
rv0	f	6/1/67	12/1/90	India	20	5
sn0	f	12/6/66	10/14/91	India	20	5
sr0	f	10/13/61	12/1/90	India	20	5
sr1	f	6/6/65	1/31/91	India	20	5
su0	m	1/31/61	3/4/90	India	20	5
vr0	m	11/7/55	12/1/90	India	20	5
vvg0	m	1/3/65	6/1/91	India	20	5
ykm0	m	2/24/65	3/20/92	India	20	0

Table B.6: OGI_HQ Segmenter Training Set

English		*Japanese*		*Mandarin*		*Tamil*	
Spkr	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
fbbf0	s1-s5	fcy0	s1-s5	ffl0	s1-s5	fhs0	s1-s5
fjw1	s1-s5	fho0	s1-s5	fjw0	q1 q2 s1-s3	fjs0	s1-s5
fka0	s1-s5	fmm0	s1-s5	fll0	s1 s10-s13	fjv0	s1-s5
fmm1	s1-s5	fks0	s1-s5	fqx0	s1-s5	fng0	s1-s5
fsjs0	s1-s5	fnu0	s1-s4 nm	fyp0	s1-s5	fsr0	s1-s5
mdcd0	s1-s5	mft0	s1-s5	mjq0	dw nm q1 q2 s3	mdr0	s1 s16-s19
mkdr0	s1-s5	mfy0	s1 s2 s16-s18	mls0	s1-s5	mmg0	s1-s5
mnjh0	s1-s5	mhw0	s1-s5	mlz0	s1-s3 s10 s11	mrps0	s1 s16-s19
mpmb0	s1-s5	mku0	q1 q2 nm s4 s5	mws0	s1 s4 s5 s10 s11	mrr0	nm s1-s4
mslb0	s1-s5	mta0	s1-s5	myt0	nm s1-s3 s5	msu0	q2 s1 s2 s4 s5

Table B.7: OGI_HQ Segmenter Development Test Set

English		*Japanese*		*Mandarin*		*Tamil*	
Spkr	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
flvl0	s1-s5	fea0	s1-s5	fxh0	s1-s5	frv0	s1-s5
mjjs0	s1-s5	mmi0	s1-s5	mcd0	s1-s5	mvr0	s1-s5

Table B.8: OGI_HQ Segmenter Final Test Set

English		*Japanese*		*Mandarin*		*Tamil*	
Spkr	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
fcp0	s1-s5	fmt0	s1-s5	fjh0	s1-s5	fsn0	s1-s5
fss0	s1-s5	fyi0	s1-s5	flz0	s1-s5	fsr1	s1-s5
mdcb0	s1-s5	mst0	s1-s5	mgl0	s1-s5	makg0	dw s1-s4
mrms0	s1-s5	mtf0	s1-s5	mny0	s1-s5	mvvg0	s1-s5

Table B.9: OGI_HQ Classifier Training Set

English	*Japanese*	*Mandarin*	*Tamil*
fbbf0	fcy0	ffl0	fhs0
fjbc0	ffi0	fjw0	fjs0
fjw1	fho0	fll0	fjv0
fka0	fmm0	fqx0	fla0
fmm1	fmt0	fsg0	fng0
fsjs0	fnu0	fyp0	fsr0
mdcd0	fyi0	mgl0	mdr0
mkdr0	mft0	mls0	mmg0
mnjh0	mfy0	mlz0	mnm0
mpmb0	mhw0	mws0	mrps0
mslb0	mku0	mxw0	mrr0
mtkl0	mta0	myt0	msu0

Table B.10: OGI_HQ Classifier Development Test Set

English	*Japanese*	*Mandarin*	*Tamil*
fcp0	fea0	fxh0	frv0
mdcb0	mmi0	mcd0	mvr0

Table B.11: OGI_HQ Classifier Final Test Set

English	*Japanese*	*Mandarin*	*Tamil*
fkw0	fmk0	fgg0	fsn0
flvl0	fmn0	fjh0	fsr1
fss0	fst0	flz0	makg0
mjjs0	mst0	myl0	mrs0
mrms0	mtf0	mzb0	mvvg0
mvcw0	mtt0	mzl0	mykm

B.3 Ten-language Telephone Speech Corpus

B.3.1 Speaker and Utterance Information

Since the callers were not asked information about their origin or date of birth, the speaker information for this corpus is limited to subjective judgments about the caller's age (adult, child) and gender (male, female, unknown). Table B.12 displays the distribution of utterances in each language by type. The utterance types are represented by the two-letter codes defined in Appendix A.

B.3.2 Segmenter Training, Development and Final Test Sets

Tables B.13 and B.14, B.15 and B.16, and B.17 and B.18 display the actual calls and utterances used in the training, development test and final test sets, respectively. The gender of the caller is juxtaposed to each call number.

B.3.3 Classifier Training, Development and Final Test Sets

Since a variable number of utterances were used from each language in the language classification data sets, Table B.19 displays the distribution of utterances from each language in the training, development test and final test sets. Tables B.20 through B.26, B.27 through B.30, and B.31 through B.34 display the callers and utterances used from each language in the training, development test and final test sets, respectively.

Table B.12: Ten-language Telephone Speech Corpus: Speaker and Utt. Information

Language	*#Calls*	*#Utts*	*Utterance Types*									
			nl	*cl*	*dw*	*nm*	*hl*	*hc*	*rm*	*ml*	*sb*	*sa*
English	868	7991	820	837	847	842	802	805	801	800	753	684
Farsi	115	993	108	108	113	113	101	102	96	93	95	64
French	122	1082	118	117	117	116	111	107	104	106	104	82
German	118	1059	110	109	116	116	105	105	105	104	101	88
Japanese	107	930	99	102	98	104	92	90	92	92	85	76
Korean	112	905	92	96	107	110	92	96	95	91	82	44
Mandarin	141	1103	130	130	120	132	102	110	108	107	98	66
Spanish	128	1150	120	123	123	125	118	118	112	117	109	85
Tamil	149	1189	143	138	133	137	126	114	112	114	98	74
Vietnamese	127	1023	115	119	121	122	94	107	104	103	91	47

Table B.13: OGI_TS Segmenter Training Set

English		*Farsi*		*French*		*German*		*Japanese*	
Spkr	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
4m	nl rm	2m	nl nm	1m	nl nm	1f	nl nm	1m	nl nm
6f	nl nm	5m	nl nm	2m	nl nm	2m	nl nm	2f	nl nm
8m	nl nm	6m	nl nm	4m	nl nm	4f	hl nm	3m	nl nm
9m	nl nm	8m	nl nm	5m	nl nm	5f	hl nm	4m	nl nm
11f	nl nm	9f	nl nm	6m	nl nm	6f	nl nm	6m	nl nm
12f	nl nm	10m	hl nl	7m	hl nl	7m	hl nl	7m	hl nl
13m	nl nm	11m	nl nm	9m	hl nl	9f	hl nl	13m	hl nl
17m	nl nm	13m	hc hl	11f	hl nl	10f	hl nl	15m	hl nl
18m	nl nm	16m	hl rm	13m	hl nl	13f	hl nl	17f	hl nl
19f	nl nm	17m	hl nl	15m	hl nl	14m	hc nl	19f	hl nl
22m	cl hl	20f	hl nl	16f	hl nl	15f	hl nl	22m	hl nl
23f	cl hl	22m	hl nl	17m	hl nl	16f	hl nl	23m	hl nl
24m	hl nl	24f	hl nl	18m	hl nl	18m	nl ml	24m	hl nl
25m	hl nl	26m	hl nl	20m	hl nl	19f	nl hl	25f	hl nl
26m	hl nl	27m	hl nl	21m	hl nl	22f	hl nl	26f	hl nl

Table B.14: OGI_TS Segmenter Training Set (continued)

Korean		*Mandarin*		*Spanish*		*Tamil*		*Vietnamese*	
Spkr	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
1f	nl nm	1m	nl nm	1m	nl nm	1m	nl nm	1m	nl nm
3f	hl nm	3f	nl nm	2f	nl nm	3m	nl nm	2f	nl nm
4m	nl nm	8m	nl nm	3m	nl nm	4m	nl nm	3m	nl nm
5m	hc rm	9m	nl nm	4m	nl nm	7m	hl nl	6m	hc nl
7f	hl nl	10m	nl nm	5m	nl nm	8m	hl nl	8f	hl nl
8m	nl nm	11m	hl nl	6m	hl nl	9m	hl nl	9m	nl nm
9m	hl nl	12m	hl nl	7m	dw nl	10f	hc nl	11m	hl nl
10m	nl nm	13m	hl nl	8m	hl nl	11m	hl nl	12m	hl nm
14m	hl nl	14f	hl nl	9m	hl nl	12m	hl nl	13m	hl nl
15m	hl nl	15m	hl nl	10f	nl nm	14m	hl nl	14f	hl nl
16f	hl nl	16f	hl nl	12m	cl hl	15m	hl nl	16m	hl nl
17m	hl nl	18m	hl nl	13m	hl nl	16m	hl nl	18f	hl nl
18m	hl nl	19m	hl nl	15f	hl nl	17m	hl nl	21f	hl nl
19m	hl nl	21m	hl nl	16f	hl nl	18m	hl nl	25f	hl nl
20m	hl nl	23m	hl nl	17m	hl nl	20m	hl nl	26m	hl nl

Table B.15: OGI_TS Segmenter Development Test Set

English		*Farsi*		*French*		*German*		*Japanese*	
Spkr	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
27m	hl rm	29m	hl ml	22f	hl ml	23f	hl ml	27m	hl ml
28m	hl ml	30m	hl ml	23m	hl ml	24m	hl ml	28m	hl nl
29m	hl ml	31m	hl ml	25f	hl ml	26m	hl ml	29m	hl ml
31f	hl ml	35f	hl ml	26m	hl ml	27m	hl ml	35f	hl ml
32f	hl ml	36m	hc hl	30m	hl ml	28m	hl ml	36m	hl ml

Table B.16: OGI_TS Segmenter Development Test Set (continued)

Korean		*Mandarin*		*Spanish*		*Tamil*		*Vietnamese*	
Spkr	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
22f	hl ml	24m	hc rm	18m	hl ml	21m	hl ml	27m	hl ml
23f	hl ml	27m	hl ml	19m	hl ml	22f	dw nl	29f	hc ml
24f	hl ml	30f	hl ml	20m	hl ml	23m	hl ml	30f	ml rm
25m	hl ml	36m	hl ml	22m	hl ml	24m	hl ml	31m	ml rm
26m	hl ml	37f	hl ml	23m	hl ml	26m	hl ml	32f	hl ml

Table B.17: OGI_TS Segmenter Final Test Set

English		*Farsi*		*French*		*German*		*Japanese*	
Spkr	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
33f	hl ml	37m	hl ml	32m	hl ml	31m	hl ml	37m	hc hl
34m	hl ml	39f	hl ml	33m	hl ml	32m	hl ml	38m	hl ml
35m	hl ml	40m	hl ml	34m	hl ml	33m	hl ml	40m	hl ml
37m	hl ml	42m	hl ml	35f	hl ml	34m	hl rm	46f	hl rm
38m	hl ml	43m	hl ml	37m	hl ml	36f	hc rm	47f	ml rm

Table B.18: OGI_TS Segmenter Final Test Set (continued)

Korean		*Mandarin*		*Spanish*		*Tamil*		*Vietnamese*	
Spkr	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
33f	hl ml	37m	hl ml	32m	hl ml	31m	hl ml	37m	hc hl
34m	hl ml	39f	hl ml	33m	hl ml	32m	hl ml	38m	hl ml
35m	hl ml	40m	hl ml	34m	hl ml	33m	hl ml	40m	hl ml
37m	hl ml	42m	hl ml	35f	hl ml	34m	hl rm	46f	hl rm
38m	hl ml	43m	hl ml	37m	hl ml	36f	hc rm	47f	ml rm

Table B.19: Distribution of Utterances in the Language Classification Data Sets

Language	*Training*	*Dev. Test*	*Final Test*
English	279	115	115
Farsi	264	111	111
French	284	115	109
German	283	118	114
Japanese	278	112	113
Korean	257	109	95
Mandarin	269	109	109
Spanish	284	111	109
Tamil	263	113	101
TOTAL	2714	1120	1077

Table B.20: OGI_TS Classifier Training Set (50 speakers/language)

	English		*Farsi*		*French*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
1.	3m	hc hl rm sa sb	5m	hc hl ml rm sa sb	1m	hc hl ml rm sa sb
2.	4m	hc hl ml rm sa sb	6m	hc hl rm	2m	hc hl ml rm sb
3.	5m	hc hl ml rm sa sb	8m	hc hl ml rm sb	5m	hc hl ml rm sa sb
4.	6f	hc hl ml rm sa sb	9f	hc hl ml rm sa sb	6m	hc hl ml rm sa sb
5.	8m	hc hl ml rm sa sb	10m	hc hl ml rm sb	7m	hc hl ml rm sb
6.	9m	hc hl ml rm sa sb	12m	hc hl ml rm sa sb	9m	hc hl ml rm sb
7.	11f	hc hl ml rm sa sb	16m	hl rm sb	11f	hc hl ml rm sa sb
8.	12f	hc hl ml rm sa sb	17m	hc hl rm sb	12m	hc hl ml rm sa sb
9.	13m	hc hl ml rm sa sb	20f	hc hl ml rm sb	13m	hc hl ml rm sa sb
10.	17m	hc hl ml rm sa	22m	hc hl ml rm sa	15m	hc hl ml rm sa sb
11.	18m	hc hl rm sa sb	24f	hc hl ml rm sa sb	16f	hc hl ml rm sa sb
12.	19f	hc hl ml sa sb	26m	hc hl ml rm sb	17m	hc hl ml rm sa sb
13.	20f	hc hl ml rm sa sb	27m	hc hl ml rm sa sb	18m	hc hl ml rm sa sb
14.	21m	hl ml sa	28f	hc hl ml rm sa sb	20m	hc hl ml rm sa sb
15.	22m	hl ml rm sa sb	29m	hc hl ml rm sa sb	21m	hc hl ml rm sa sb
16.	23f	hc hl ml rm	30m	hc hl ml rm sb	22f	hc hl ml rm sb
17.	24m	hc hl ml rm sa sb	31m	hc hl ml rm sb	23m	hc hl ml rm sa sb
18.	25m	hc hl ml rm sa	35f	hc hl ml rm sa sb	25f	hc hl ml rm sa sb
19.	26m	hc hl ml rm sa	36m	hc hl	26m	hc hl ml rm sa sb
20.	27m	hc hl rm sb	37m	hc hl ml rm sa sb	29m	hc hl ml rm sb
21.	28m	hc hl ml sa sb	39f	hc hl ml sb	30m	hc hl ml rm sa sb
22.	29m	hc hl ml rm sa sb	40m	hc hl ml rm sa sb	32m	hc hl ml rm sa sb
23.	30m	hl ml rm sa sb	41f	hc rm sb	33m	hc hl ml rm sa sb
24.	31f	hc hl ml rm sa sb	42m	hc hl ml rm sa sb	34m	hc hl ml rm sa sb
25.	32f	hc hl ml rm sa sb	43m	hc hl ml rm sb	35f	hc hl ml rm sa sb

Table B.21: OGI_TS Classifier Training Set (continued)

	English		*Farsi*		*French*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
26.	33f	hc hl ml rm sa sb	44m	hc hl ml rm sa sb	37m	hc hl ml rm sa sb
27.	34m	hc hl ml rm sa sb	45f	hc hl ml rm sa sb	38m	hc hl ml rm sa sb
28.	35m	hc hl ml rm sa sb	46f	hc ml rm	40m	hc hl ml rm sa sb
29.	37m	hc hl ml rm sb	47m	hc hl ml rm sa sb	41f	hc hl ml rm sa sb
30.	38m	hc hl ml rm sa sb	51m	hc hl ml rm sa sb	42f	hc hl ml rm sa sb
31.	39f	hc hl ml rm	52m	hc hl ml rm sa sb	43m	hc hl
32.	40m	hc hl ml rm sa sb	53m	hc hl ml rm sa sb	44m	hc hl ml rm sa sb
33.	41f	hc hl ml rm sa sb	55m	hc hl ml rm sa sb	45f	hc hl ml rm sa sb
34.	42m	hc hl ml rm sa sb	56m	hc hl ml rm sa sb	46m	hc hl ml rm sa sb
35.	43m	hc hl ml rm sa sb	59m	hc hl ml rm sb	47m	hc hl ml rm
36.	44m	hc hl ml rm sa sb	62f	hc hl ml rm sa sb	48m	hc hl ml rm sa sb
37.	45m	hc hl ml rm sa sb	63m	hc hl ml rm sa sb	51f	hc hl ml rm sa sb
38.	47f	hc hl ml rm sa sb	66m	hc hl ml rm sa sb	52m	hc hl ml rm sa sb
39.	48f	hc hl ml rm sb	67m	hc hl ml rm sb	53m	hc hl ml rm sb
40.	50f	hc hl ml rm sa sb	70m	hc hl ml rm sa sb	55m	hc hl ml rm sa sb
41.	51f	hc hl ml rm sa sb	72m	hc hl ml rm sa sb	56m	hc hl ml rm sa sb
42.	52m	hc hl ml rm sa sb	74m	hc hl ml rm sa sb	57m	hc hl
43.	53m	hc hl ml rm sa sb	77m	hc hl rm sa sb	58f	hc hl ml rm sa sb
44.	54m	hc hl ml rm sa sb	78m	hc hl ml rm sa sb	59m	hc hl ml rm sa sb
45.	56f	hc hl ml rm sa sb	82m	hc hl rm sa sb	61m	hc hl ml rm sa sb
46.	57m	hc hl ml rm sa sb	83f	hc hl ml rm sb	62m	hc hl ml rm sa sb
47.	58m	hc hl ml rm sa sb	84m	hc hl ml rm sb	64m	hc hl ml rm sa sb
48.	59m	hc ml rm sa sb	85m	hc hl ml rm sb	65m	hc hl ml rm sa sb
49.	60m	hc hl ml rm sa sb	86m	hc hl ml rm sb	66m	hc hl ml rm sa sb
50.	61m	hc hl ml rm sa sb	88m	hc hl ml rm sa sb	67m	hc hl ml rm sa sb

Table B.22: OGI_TS Classifier Training Set (continued)

	German		*Japanese*		*Korean*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
1.	1f	hc hl ml rm sa sb	1m	hc hl ml rm sa sb	1f	hc hl ml rm
2.	2m	hc hl ml rm sa sb	2f	hc hl ml rm sa sb	3f	hc hl ml rm sa sb
3.	3m	hc hl ml rm sa sb	3m	hc hl ml rm sa sb	4m	hc hl ml rm sa sb
4.	4f	hc hl ml rm sa sb	4m	hc ml	7f	hc hl ml rm sb
5.	5f	hc hl ml rm sa sb	6m	hc hl ml rm	9m	hc hl ml rm sb
6.	6f	hc hl ml rm sa sb	7m	hc hl ml rm sa sb	14m	hc hl ml rm sb
7.	7m	hc hl ml rm sa sb	13m	hc hl ml rm sa sb	15m	hc hl ml rm sa sb
8.	9f	hc hl ml rm sa sb	15m	hc hl ml rm sa sb	16f	hc hl ml rm
9.	10f	hc hl ml rm sa sb	17f	hc hl ml rm sa sb	17m	hc hl ml rm sa sb
10.	11f	hc hl ml rm sa sb	19f	hc hl ml rm sa sb	18m	hc hl ml rm sa sb
11.	12f	hc hl ml rm sb	20m	hc hl ml rm sa sb	19m	hc hl ml rm sa sb
12.	13f	hc hl ml rm	22m	hc hl ml rm sa sb	20m	hc hl ml rm
13.	14m	hc sa sb	23m	hc hl ml rm sa sb	22f	hc hl ml rm sa sb
14.	15f	hc hl ml rm sb	24m	hc hl ml rm sa sb	23f	hc hl ml rm sb
15.	16f	hc hl ml rm sa sb	25f	hc hl rm sb	24f	hc hl ml rm sb
16.	18m	hc hl ml rm sa sb	26f	hl ml rm sa sb	25m	hc hl ml rm sb
17.	19f	hc hl ml rm sa sb	27m	hl ml rm sb	26m	hc hl ml rm sa sb
18.	22f	hc hl ml rm sa sb	29m	hc hl ml rm sa sb	30f	hc hl ml rm sa sb
19.	23f	hc hl ml rm sa sb	35f	hc hl ml rm sa sb	34m	hc ml rm sb
20.	24m	hc hl ml rm sa sb	36m	hc hl ml rm sa sb	35m	hc hl ml rm
21.	26m	hc hl ml rm sa sb	37m	hc hl	36m	hc hl ml rm sb
22.	27m	hc hl ml rm sa sb	38m	hc hl ml rm sa sb	37m	hc hl ml rm sb
23.	28m	hc hl ml rm sa sb	40m	hc hl ml rm sa sb	38m	hc ml rm sb
24.	31m	hc hl ml rm sa sb	47f	hc ml rm sb	39m	hc hl ml rm sb
25.	32m	hc hl ml rm	48m	hc hl ml rm sa sb	40m	hc hl ml rm sb

Table B.23: OGI_TS Classifier Training Set (continued)

	German		*Japanese*		*Korean*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
26.	33m	hc hl ml rm sa sb	50f	hc hl ml rm sa sb	41m	hc hl ml rm sa sb
27.	34m	hc hl rm sa sb	51m	hc hl ml rm sa sb	42m	hc hl ml rm sa sb
28.	36f	hc hl ml rm sa sb	53f	hc hl ml rm sa sb	43m	hc hl ml rm
29.	37m	hc hl ml rm sa sb	54m	hl ml rm sa sb	46f	hc hl ml rm sa sb
30.	38m	hc hl ml rm sa sb	55f	hc hl ml rm sa sb	48f	hc hl ml rm sb
31.	39m	hc hl ml rm sa sb	57m	hc hl ml rm sa sb	49m	hc hl ml rm sb
32.	40m	hc hl ml rm sb	58m	hc hl ml rm sa sb	51f	hc hl
33.	41m	hc hl ml rm sa sb	60f	hc hl ml rm sa sb	53m	hc hl ml rm sa sb
34.	42f	hc hl ml rm sa sb	61m	hc hl ml rm sa sb	54m	hc hl ml rm sa sb
35.	44f	hc hl ml rm sa sb	62m	hc ml rm sb	55m	hc hl ml rm sb
36.	45f	hc hl ml rm sa sb	65m	hc hl ml rm sa sb	56m	hc hl ml rm sb
37.	46f	hc hl ml rm sa sb	66m	hc hl ml rm sa sb	58m	hc hl ml rm sb
38.	47m	hc hl ml rm sa sb	67f	hc hl ml rm sa sb	59f	hc hl ml rm sa sb
39.	50f	hc hl ml rm sa sb	68f	hc hl ml rm sa sb	60m	hc hl ml rm sb
40.	51m	hc hl ml rm sa sb	69m	hc hl ml rm sb	61m	hc hl ml rm sa sb
41.	52m	hl ml rm	71f	hc hl ml rm sa sb	62m	hc hl ml rm sb
42.	53f	hc hl ml rm sa sb	72f	hc hl ml rm sa sb	63f	hc hl ml rm sa
43.	56f	hc hl ml rm sa sb	73f	hc hl ml rm sa sb	64m	hc hl ml rm sb
44.	57f	hc hl ml rm sa sb	75m	hc hl ml rm sb	65m	hc hl ml rm
45.	58m	hc hl ml rm sb	80m	hc hl ml rm sa sb	67f	hc hl ml rm sb
46.	59f	hc hl ml rm sa sb	82f	hc hl ml rm sa sb	68m	hc hl ml rm sa sb
47.	60m	hc hl ml rm sa sb	83m	hc hl ml rm sa sb	69f	hc hl ml rm sa sb
48.	61m	hc hl ml rm sa sb	85f	hc hl ml rm sa sb	70m	hc hl ml rm sa sb
49.	63f	hc hl ml rm sa sb	86f	hc hl ml rm sa sb	71f	hc hl ml rm sa sb
50.	68m	hc hl ml rm	88m	hc hl ml rm sa sb	72m	ml rm sb

Table B.24: OGI_TS Classifier Training Set (continued)

	Mandarin		*Spanish*		*Tamil*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
1.	1m	hc hl ml rm sb	1m	hc hl ml rm sa sb	1m	hc hl ml rm sa sb
2.	8m	hc ml rm	2f	hc hl ml rm sa sb	3m	hc hl ml rm sb
3.	9m	hc hl ml rm sa sb	3m	hc hl ml rm sb	4m	hc hl ml rm sa sb
4.	10m	hc hl ml rm	4m	hc hl ml rm sb	7m	hl sb
5.	11m	hc hl ml rm sa sb	5m	hc hl ml rm sa sb	8m	hc hl ml rm sa sb
6.	12m	hc hl ml rm	6m	hc hl ml rm sa sb	9m	hc hl ml rm sa sb
7.	13m	hc hl ml rm sa sb	8m	hc hl ml rm sa sb	11m	hc hl ml rm sa sb
8.	14f	hc hl ml rm sa sb	9m	hc hl ml rm	12m	hc hl ml rm sa sb
9.	15m	hc hl ml rm sa sb	10f	hc ml rm sa sb	14m	hc hl ml rm sa sb
10.	16f	hc hl ml rm sa sb	12m	hc hl ml rm sa sb	15m	hc hl ml rm sa sb
11.	18m	hc hl ml rm sa sb	13m	hc hl ml rm sa sb	16m	hc hl ml rm sa sb
12.	21m	hc hl ml rm sa sb	14m	hc hl ml rm sa sb	17m	hc hl ml rm
13.	23m	hc hl ml rm sb	15f	hc hl ml rm sa sb	18m	hc hl ml rm sa sb
14.	24m	hc ml rm sb	16f	hc hl ml rm sa sb	21m	hc hl ml rm sb
15.	27m	hc hl ml rm sa sb	17m	hc hl ml rm sb	23m	hl ml rm
16.	30f	hc hl ml rm sa sb	18m	hc hl ml rm sa sb	24m	hc hl ml rm sa sb
17.	31f	hc hl ml rm sa sb	19m	hc hl ml rm sa sb	26m	hc hl ml rm sb
18.	32f	hc hl	20m	hc hl ml rm sa sb	27m	hl ml rm sa sb
19.	33f	hc hl ml rm sa sb	22m	hc hl ml rm sa sb	29f	hc hl rm
20.	34m	hc hl ml rm sb	23m	hc hl ml rm sb	30m	hc hl ml sb
21.	35m	hc hl ml rm sb	24m	hc hl ml rm sb	32m	hc hl ml rm sb
22.	36m	hc hl ml rm sa sb	25f	hc hl ml rm sa sb	33m	hc hl ml rm sa sb
23.	37f	hc hl ml rm sa sb	26m	hc hl ml rm sa sb	34m	hc hl ml rm
24.	39m	hc hl ml rm sa sb	27m	hc hl ml rm sa sb	35m	hc hl ml rm sb
25.	40f	hc hl ml rm sb	28m	hc hl ml rm sa sb	36m	hc hl ml rm

Table B.25: OGI_TS Classifier Training Set (continued)

	Mandarin		*Spanish*		*Tamil*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
26.	41m	hc hl ml rm sa sb	29m	hc hl ml rm sa sb	37m	hc hl ml rm
27.	42f	hc hl ml rm sb	30f	hc hl ml rm sa sb	38m	hc hl ml rm sa sb
28.	43m	hc hl ml rm sa sb	31m	hc hl ml rm sa sb	39m	hc hl ml rm sa sb
29.	44f	hc hl ml rm sa sb	32f	hc hl ml rm sa sb	40m	hc hl ml rm sa sb
30.	46m	hc hl ml rm sa sb	33f	hc hl ml rm sa sb	41m	hc hl ml rm sa sb
31.	48m	hc hl ml rm sa sb	35f	hc hl ml rm sa sb	43m	hc hl ml rm sa sb
32.	49m	hc hl ml rm sa sb	36m	hc hl ml rm sa sb	44m	hc hl ml rm sa sb
33.	51f	hc hl ml rm sa sb	37f	hc hl ml rm sa sb	45f	hc hl ml rm sa sb
34.	52m	hc hl ml rm sb	38f	hc hl ml rm sa sb	48f	hc ml
35.	53m	hc hl ml rm sb	39m	hc hl	50m	hc hl ml rm sb
36.	55m	hc ml rm sb	40f	hc hl ml sb	51m	hc hl ml rm sa sb
37.	56m	hc hl ml rm sa sb	41m	hc hl ml rm sa sb	56m	hc hl ml rm sa sb
38.	57m	hc hl ml rm sb	44f	hc hl ml rm sa sb	57m	hc hl ml rm sa sb
39.	58m	hc hl ml rm sa sb	46m	hc hl ml rm sb	59m	hc hl ml rm sa sb
40.	59f	hc hl ml rm sa sb	47f	hc hl ml rm sa sb	60m	hc hl ml rm sa sb
41.	60m	hc hl ml rm sb	48m	hc hl ml rm sa sb	61f	hc hl ml rm sa sb
42.	61m	hc hl rm	49m	hc hl ml rm sa sb	63f	hc hl ml rm sb
43.	65f	hc hl ml rm sb	50m	hc hl ml rm sa sb	65m	hc hl ml rm sa sb
44.	67f	hc hl ml rm sa sb	51m	hc hl ml rm sa sb	66m	hc hl ml sb
45.	68f	hc hl ml rm sa sb	52f	hc hl ml rm sa sb	67m	hc hl ml rm sa sb
46.	69m	hc hl ml rm sa sb	53m	hc hl ml rm sa sb	68m	hc hl ml rm sa sb
47.	73f	hc hl ml rm sb	54m	hc hl ml rm sa sb	69m	hl rm sb
48.	76m	hc hl ml rm sa sb	55m	hc hl ml rm sb	71f	hc hl ml rm sa sb
49.	77m	hc hl ml rm sa sb	56f	hc hl ml rm sa sb	73m	hc hl ml rm sa sb
50.	78m	hc hl ml rm sb	57m	hc hl ml rm sa sb	74m	hc hl ml rm sa sb

Table B.26: OGI_TS Classifier Training Set (continued)

Vietnamese					
No.	*Spkr*	*Utts.*	*No.*	*Spkr*	*Utts.*
1.	1m	hc hl ml rm sa sb	26.	41f	hc hl ml rm sb
2.	2f	hc hl ml rm sb	27.	42m	hc hl ml rm sa sb
3.	3m	hc hl ml rm	28.	43f	hc hl ml rm sa sb
4.	8f	hc hl ml rm sb	29.	44f	hc hl ml rm sa sb
5.	11m	hc hl ml rm sa sb	30.	45f	hc hl ml rm sa sb
6.	12m	hc hl ml rm sa sb	31.	46m	hc hl ml rm sb
7.	13m	hc hl ml rm sb	32.	47m	hc hl ml rm sa sb
8.	14f	hc hl ml rm sa sb	33.	49f	hc hl ml rm
9.	16m	hc hl	34.	52m	hc hl ml rm sa sb
10.	18f	hc hl ml rm sb	35.	53m	hc hl ml rm sb
11.	21f	hc hl ml rm sb	36.	55m	hc hl ml rm sa sb
12.	25f	hc hl ml rm sb	37.	56m	hc hl ml rm sa sb
13.	26m	hc hl ml rm sb	38.	57m	hc ml rm sa sb
14.	27m	hc hl ml rm sa sb	39.	58m	hc hl ml rm sb
15.	29f	hc ml rm	40.	63m	hc hl ml rm sa sb
16.	30f	hc ml rm	41.	64m	hc hl ml rm sa sb
17.	31m	ml rm	42.	65m	hc hl ml rm sa sb
18.	32f	hc hl ml rm	43.	68f	hc hl ml rm sa sb
19.	33m	hc hl ml rm sb	44.	69m	hc hl ml rm sb
20.	34f	hc hl ml rm sb	45.	70f	hc hl ml rm sb
21.	35f	hc hl ml rm sb	46.	71m	hc hl ml rm sa sb
22.	37m	hc hl ml rm sa sb	47.	72f	hc hl ml rm
23.	38m	hc hl ml rm	48.	74m	hc hl rm
24.	39m	hc hl ml rm sb	49.	75m	hc hl ml rm sa sb
25.	40m	hc hl ml rm sa sb	50.	76m	hc ml rm sb

Table B.27: OGI_TS Classifier Development Test Set (20 speakers/language)

	English		*Farsi*		*French*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
1.	84m	hc hl ml rm sa sb	117m	hc hl ml rm sb	97f	hc hl ml rm sa sb
2.	85m	hc hl ml rm	118f	hc hl ml rm sa sb	98f	hc hl ml rm sa sb
3.	86m	hc hl ml rm sa sb	119m	hc hl ml rm sb	99m	hc hl ml rm sa sb
4.	87m	hc hl ml rm sa sb	120m	hc hl ml rm sa sb	100f	hc hl ml rm sa sb
5.	88f	hc hl ml rm sa sb	121m	hc hl ml rm sa sb	101m	hc hl ml rm sa sb
6.	90f	hc hl ml rm sa sb	122m	hc hl ml rm sa sb	102m	hc hl ml rm sa sb
7.	92m	hc hl ml rm sa sb	123m	hc hl ml rm sa sb	103m	hc hl ml rm sa sb
8.	93m	hc hl ml rm sa sb	124f	hc hl ml rm sa sb	104f	hc hl ml rm sb
9.	94m	hc hl ml rm sa sb	125m	hc hl	105m	hc hl ml rm sa sb
10.	96m	hc hl ml rm sa sb	126m	hc hl ml rm sa sb	106m	hc hl ml rm sb
11.	97m	hc hl ml rm sa sb	127m	hc hl ml rm sb	107m	hc hl ml rm sa sb
12.	98m	hc hl ml rm sa sb	128f	hc hl ml rm sa sb	108m	hc hl ml rm sa sb
13.	99f	hc hl ml rm sa sb	129m	hc hl ml rm sa sb	109m	hc hl ml rm sb
14.	100m	hc hl ml rm sa sb	130m	hc hl ml rm sa sb	110m	hc hl ml rm sa sb
15.	101m	hc hl ml rm	131m	hc hl ml rm sa sb	113m	ml rm sa sb
16.	103f	hc hl ml rm sb	132m	hc hl ml rm sa sb	114f	hc hl ml rm sa sb
17.	105m	hc hl ml rm sa sb	135m	hc hl ml rm sa sb	116f	hc hl ml rm sa sb
18.	106f	hc hl ml rm sa sb	136m	hc hl ml rm sa sb	117m	hc hl ml rm sa sb
19.	107m	hc hl ml rm sa sb	138f	hc hl ml rm sb	118m	hc hl ml rm sa sb
20.	108f	hc hl ml rm sa sb	139m	hc hl ml rm sb	119f	hc hl ml rm sa sb

Table B.28: OGI_TS Classifier Development Test Set (continued)

	German		*Japanese*		*Korean*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
1.	95m	hc hl ml rm sa sb	118m	hc hl ml rm sa sb	109m	hc hl ml rm sa sb
2.	97f	hc hl ml rm sa sb	120m	hc hl ml rm sa sb	110m	hc hl ml rm sa sb
3.	99m	hc hl ml rm sa sb	121f	hc hl ml rm sa sb	111m	hc hl ml rm sa sb
4.	100f	hc hl ml rm sa sb	122m	hc hl ml rm sa sb	112m	hc hl ml rm sa sb
5.	101m	hc hl ml rm sa sb	124f	hc hl ml rm sa sb	113m	hc hl ml rm sa sb
6.	102m	hc hl ml rm sa sb	126m	hc hl ml rm sa sb	114m	hc hl ml rm sb
7.	106m	hc hl ml rm sa sb	127m	hc ml rm sa sb	118m	hc hl ml rm sa sb
8.	109m	hc hl ml rm sa sb	128f	hc hl ml rm	120m	hc hl ml rm sa sb
9.	113f	hc hl ml rm sa sb	129m	hc hl ml rm sa sb	125m	hc hl ml rm sb
10.	114m	hc hl ml rm sa sb	130m	hc hl ml rm	127m	hc hl rm sa
11.	116f	hc hl ml rm sa sb	131m	hc hl ml rm sa sb	129m	hc hl ml rm sa sb
12.	118f	hc hl ml rm sa sb	133m	hc hl ml rm sa sb	130m	hc hl ml rm sa sb
13.	120f	hc hl ml rm sa sb	134m	hc hl ml rm	131m	hc hl ml rm sa sb
14.	123m	hc hl ml rm sa sb	135m	hc hl ml rm sa sb	132m	rm sa sb
15.	124f	hc hl ml rm sa sb	136m	hc hl ml rm sa sb	136f	hc hl ml rm sb
16.	125m	hc hl ml rm sa sb	137f	hc hl ml rm sa sb	137m	hc hl ml rm sb
17.	127m	hc hl ml rm sb	138m	hc hl ml rm sa sb	138m	hc hl ml rm sb
18.	128f	hc hl ml rm sa	139m	hc hl ml rm sa sb	139m	hc hl ml rm sa sb
19.	129f	hc hl ml rm sa sb	140m	hc hl ml rm sb	140f	hc hl ml rm sa
20.	130m	hc hl ml rm sa sb	141m	hc hl ml rm sa sb	141m	hc hl ml rm sa sb

Table B.29: OGI_TS Classifier Development Test Set (continued)

	Mandarin		*Spanish*		*Tamil*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
1.	121f	hc hl ml rm sa sb	81m	hc hl ml rm sa sb	105m	hc hl ml rm sa sb
2.	122f	hc hl ml rm sa sb	82m	hc hl ml rm sa sb	106m	hc hl ml rm sa sb
3.	123f	hc hl ml rm sa sb	83m	hc hl ml rm sa sb	107m	hc hl ml rm sa sb
4.	124f	hc hl ml rm sb	84m	hc hl ml rm sa sb	108m	hc hl ml rm sa sb
5.	126m	hc hl ml rm sb	85m	hc hl ml rm sb	110m	hc hl ml rm sa sb
6.	127m	hc hl ml rm sb	87m	hc hl ml rm sa sb	111m	hc hl ml rm sa sb
7.	129f	hc hl ml rm sa sb	88m	hc hl ml rm sb	114m	hc hl ml rm sb
8.	134m	hc hl ml rm sb	89m	hc hl ml rm sa sb	115m	hc hl ml rm
9.	135m	hc hl ml rm sa sb	90m	hc hl ml rm sa sb	116m	hc hl ml rm sa sb
10.	136m	hc hl ml rm sb	91m	hc hl ml rm sb	117f	hc hl ml rm sb
11.	137f	hc hl ml rm sa sb	92m	hc hl ml rm	118m	hc hl ml rm sb
12.	138m	hc hl ml rm sa sb	93m	hc hl ml rm sb	121m	hc hl ml rm sa sb
13.	140m	hc ml rm sb	94f	hc hl ml rm sb	122m	hc hl ml rm sa sb
14.	141m	hc hl ml rm	95m	hc hl ml rm sa sb	124m	hc hl ml rm sa sb
15.	142m	hc hl ml rm sa sb	96f	hc hl ml rm sa sb	125m	hc hl ml rm sa sb
16.	143f	hc ml rm sa sb	97m	hc hl ml rm sb	129m	hc hl ml rm sa sb
17.	146m	hc hl ml rm sa sb	98m	hc hl ml rm sa sb	132m	hc hl ml rm sa sb
18.	147m	hc hl ml rm sa sb	99f	hc hl ml rm sa sb	133f	hc hl ml rm sa sb
19.	148m	hc hl ml rm sa sb	100m	hc hl ml rm sb	137m	hc hl ml sb
20.	149m	hc hl ml rm sb	102f	hc hl ml rm sa sb	138m	hc hl ml rm sa sb

Table B.30: OGI_TS Classifier Development Test Set (continued)

Vietnamese					
No.	*Spkr*	*Utts.*	*No.*	*Spkr*	*Utts.*
1.	107f	hc hl ml rm sb	11.	126m	hc hl ml rm sb
2.	108f	hc hl ml rm	12.	127m	hc hl ml rm sa sb
3.	112m	hc hl ml rm sa sb	13.	129m	hc hl ml rm sa sb
4.	113m	hc hl ml rm sa sb	14.	130m	hc hl ml rm sb
5.	114f	hc hl ml rm sa sb	15.	131m	hc hl ml rm sb
6.	116m	hc hl ml rm sb	16.	132m	hc hl ml rm
7.	117m	hc hl ml rm sa sb	17.	134m	hc hl ml rm sb
8.	118f	hc hl ml rm sa sb	18.	135m	hc ml rm sa sb
9.	119m	hc hl ml rm sa sb	19.	139m	hc hl ml rm sa sb
10.	120m	hc hl ml rm sb	20.	140m	hc hl ml rm sb

Table B.31: OGI_TS Classifier Final Test Set (20 speakers/language)

	English		*Farsi*		*French*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
1.	62m	hc hl ml rm sa sb	90m	hc hl ml rm sa sb	68m	hc hl ml rm sa sb
2.	63m	hc hl ml rm sa sb	93f	hc hl ml rm sa sb	69m	hc hl ml rm sb
3.	64m	hc hl ml rm sa sb	94m	hc hl ml rm sa sb	71m	hc hl ml rm sa sb
4.	65m	hc rm sa sb	95m	hc hl ml rm sb	72m	hc hl ml rm sa sb
5.	66m	hc hl ml rm sa sb	96m	hc hl ml rm sb	73m	hc hl ml rm sa sb
6.	68f	hc hl ml rm sa sb	99m	hc hl ml rm sa sb	74m	hc hl ml rm sa sb
7.	69m	hc hl ml rm sa sb	100m	hc hl ml rm sa sb	75m	hc hl ml rm sa sb
8.	70m	hc hl ml rm sa sb	101m	hc hl ml rm sa sb	79f	hc hl
9.	71m	hc hl ml rm sa sb	102m	hc hl ml rm sa sb	80f	hc hl ml rm sa sb
10.	72m	hc hl ml rm sa sb	103m	hc hl ml rm sa sb	81f	hc hl ml rm sb
11.	73m	hc hl ml rm sa sb	104m	hc hl ml rm sb	83m	hc hl ml rm sa sb
12.	74f	hc hl ml rm sb	107m	hc hl ml rm sa sb	85m	hc hl ml rm sa sb
13.	76m	hc hl ml rm sa sb	108m	hc hl ml rm sa sb	86f	hc hl ml rm sb
14.	77f	hc hl ml rm sa sb	109m	hc hl ml rm sa sb	87f	hc hl ml rm sb
15.	78m	hc hl ml rm sa sb	111m	hc hl ml rm sb	90	hc hl ml rm sa sb
16.	79m	hc hl ml rm sa sb	112f	hc hl ml rm sb	92f	hc hl ml rm sb
17.	80m	hc hl ml rm	113m	hc hl ml rm sa sb	93f	hc hl ml rm sa sb
18.	81m	hc hl ml rm sa sb	114m	hc hl ml sb	94f	hc hl ml rm
19.	82m	hc hl ml rm sa sb	115m	hc hl ml rm sb	95m	hc hl ml rm sa sb
20.	83m	hc hl ml rm sa sb	116m	hc hl ml rm sb	96m	hc hl ml rm sa sb

Table B.32: OGI_TS Classifier Final Test Set (continued)

	German		*Japanese*		*Korean*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
1.	69f	hc hl ml rm sa sb	90f	ml sa sb	73f	hc ml rm sb
2.	70m	hc hl ml rm sa sb	91f	hc hl ml rm sa sb	74m	hc hl ml rm sb
3.	72m	hc hl ml rm sb	92f	hc hl ml rm sa sb	75m	hc hl ml rm sb
4.	74m	hc hl ml rm sa sb	94f	hc hl ml rm sa sb	76m	hc hl ml rm sb
5.	75f	hc hl ml rm sa sb	96m	hc hl ml rm sa sb	79m	hc hl ml rm sb
6.	77f	hc hl ml rm sa sb	97f	hc hl ml rm sa sb	80m	hc hl ml rm sb
7.	78m	hc hl ml rm sb	100m	hc hl ml rm sa sb	81m	hc hl rm sb
8.	79m	hc hl ml rm sb	101f	hc hl ml rm sa sb	85f	hc ml rm sb
9.	80m	hc hl ml rm sa sb	102m	hc hl ml rm sa sb	90m	hc hl ml rm sb
10.	81m	hc hl ml rm sa sb	104m	hc hl ml rm sb	91m	hc ml rm
11.	83m	hc hl ml rm sb	105f	hc hl ml rm sa sb	94m	hc hl ml rm sa sb
12.	85f	hc hl ml rm sb	106m	hc hl ml rm sa sb	96m	hc hl ml rm
13.	86m	hc hl ml rm sa sb	107m	hc hl ml rm sa sb	97f	hc hl ml rm sb
14.	87m	hc hl ml rm sb	108m	hc hl ml rm	99f	hc hl ml rm sa sb
15.	88m	hc hl ml rm sa sb	109f	hc hl ml rm sb	100m	hc hl ml rm sa
16.	89m	hc hl ml rm sa sb	110f	hc hl ml rm sa sb	101m	hc hl ml rm sb
17.	90m	hc hl ml rm sa sb	112m	hc hl ml rm sa sb	103f	hc hl ml rm sb
18.	91m	hc hl ml rm sa sb	113m	hc hl ml rm sa sb	105m	hc hl ml rm sb
19.	93m	hc hl ml rm sa sb	116m	hc hl ml rm sa sb	106m	hc hl ml rm sa sb
20.	94f	hc hl ml rm sa sb	117m	hc hl ml rm sa sb	108m	hc hl sb

Table B.33: OGI_TS Classifier Final Test Set (continued)

	Mandarin		*Spanish*		*Tamil*	
No.	*Spkr*	*Utts.*	*Spkr*	*Utts.*	*Spkr*	*Utts.*
1.	79m	hc hl ml rm sa sb	60m	hc hl ml rm sb	75m	hc hl rm
2.	81f	hc hl	62m	hc hl ml rm sa sb	76m	hc hl ml rm
3.	83f	hc hl ml rm sa sb	63m	hc hl ml rm sb	77m	hc hl ml rm sa sb
4.	86f	hc ml rm sa sb	64f	hc hl ml rm sb	78m	hc hl ml rm sb
5.	90f	hc hl ml rm sa sb	65m	hc hl ml rm sa sb	79m	hc hl ml rm sb
6.	92f	hc hl ml rm sa sb	66f	hc hl ml rm	80m	hc hl ml rm sa sb
7.	93m	hc hl ml rm sa sb	67m	hc hl ml sa sb	83m	hc hl ml rm
8.	97f	hc hl ml rm sa sb	68m	hc hl ml rm sa sb	84m	hc hl ml rm sa sb
9.	98f	hc hl ml rm sa sb	69f	hc hl ml rm sa sb	85m	hc hl ml rm
10.	99m	hc hl ml rm	70m	hc hl ml rm sa sb	86m	hc hl ml rm sb
11.	100f	hc hl ml rm sa sb	71f	hc hl ml rm sb	89m	hc hl ml rm sa sb
12.	101f	hc hl ml rm sa sb	72m	hc hl ml rm sa sb	90m	hc hl ml rm sa sb
13.	105m	hc hl ml rm sa sb	73m	hc hl ml rm sa sb	91m	hc hl ml rm sb
14.	106m	hc ml rm sb	74f	hc hl ml rm	94f	hc hl ml rm sb
15.	107m	hc hl ml rm sb	75m	hc hl ml rm	95m	hc hl ml rm sa sb
16.	108m	hc hl ml rm sa sb	76m	hc hl ml rm sa sb	98m	hc hl ml rm sa sb
17.	109m	hc hl ml rm sa sb	77f	hc hl ml rm sa sb	99m	hc hl ml rm
18.	113m	hc hl ml rm sa sb	78m	hc hl ml rm sa sb	100m	hl ml rm
19.	118f	hc hl ml rm sb	79f	hc hl ml rm sa sb	102m	hc hl ml rm sa sb
20.	119m	hc hl ml rm sa sb	80f	hc hl ml rm sa sb	104m	hc hl ml rm sa sb

Table B.34: OGI_TS Final Test Set (continued)

Vietnamese					
No.	*Spkr*	*Utts.*	*No.*	*Spkr*	*Utts.*
1.	77f	hc hl ml rm sb	11.	96m	hc hl rm
2.	80f	hc hl ml rm	12.	97m	hc hl ml rm sb
3.	83m	hc hl ml rm sb	13.	98m	hc ml rm sb
4.	85m	hc hl ml rm sa sb	14.	99m	hc hl ml rm sa sb
5.	86m	hc hl ml rm sa sb	15.	100m	hc hl ml rm sb
6.	87m	hc hl ml rm sa sb	16.	101m	hc hl ml rm sb
7.	89m	hc hl ml rm sa sb	17.	102f	hc ml rm sa sb
8.	90f	hc hl ml rm	18.	104f	hc hl ml rm sa sb
9.	92m	hc hl ml rm sa sb	19.	105f	hc hl ml rm sb
10.	94m	hc hl ml rm sb	20.	106f	hc hl ml rm

Appendix C

Boxplots of Feature Sets

C.1 Segment-triples

Figures C.1 through C.21 display the boxplots for the 63 segment-triple frequency (STF) features.

Figures C.22 through C.42 display the boxplots for the 63 segment-triple ratio (STR) features.

C.2 Segment-pairs

Figures C.43 through C.49 display the boxplots for the 20 segment-pair frequency (SPF) features.

Figures C.50 through C.56 display the boxplots for the 20 segment-pair ratio (SPR) features.

Figures C.57 through C.63 display the boxplots for the 20 segment-pair duration ratio (SPDR) features.

Figures C.64 through C.70 display the boxplots for the 20 segment-pair median duration (SPMD) features.

C.3 Pitch-based Features

Figures C.71 through C.73 display boxplots for the 4 intra-segment pitch variation and 4 inter-segment pitch variation features respectively.

C.4 Other Global Features

C.4.1 Frequency of Occurrence

Figures C.74 through C.77 display boxplots of the 11 frequency-of-occurrence features described in Section 5.4.9.1. OBS, VOBS, SON and SEG refer to obstruents, voiced obstruents, sonorants and all segments, respectively.

C.4.2 Segment Occurrence Ratios

Figures C.78 through C.92 display boxplots of the 47 segment occurrence ratios described in Section 5.4.9.2.

C.4.3 Segment Duration Ratios

Figures C.93 through C.107 display boxplots of the 47 segment duration ratios described in Section 5.4.9.2.

C.4.4 Duration

The duration features are described in Section 5.4.9.3. Figures C.108 through C.110 display boxplots of the 7 average duration features.

Figures C.111 through C.113 display boxplots of the 7 standard-deviation-of-duration features.

Figures C.114 through C.117 display boxplots of the 5 inter-segment duration difference and 5 vowel center distance features.

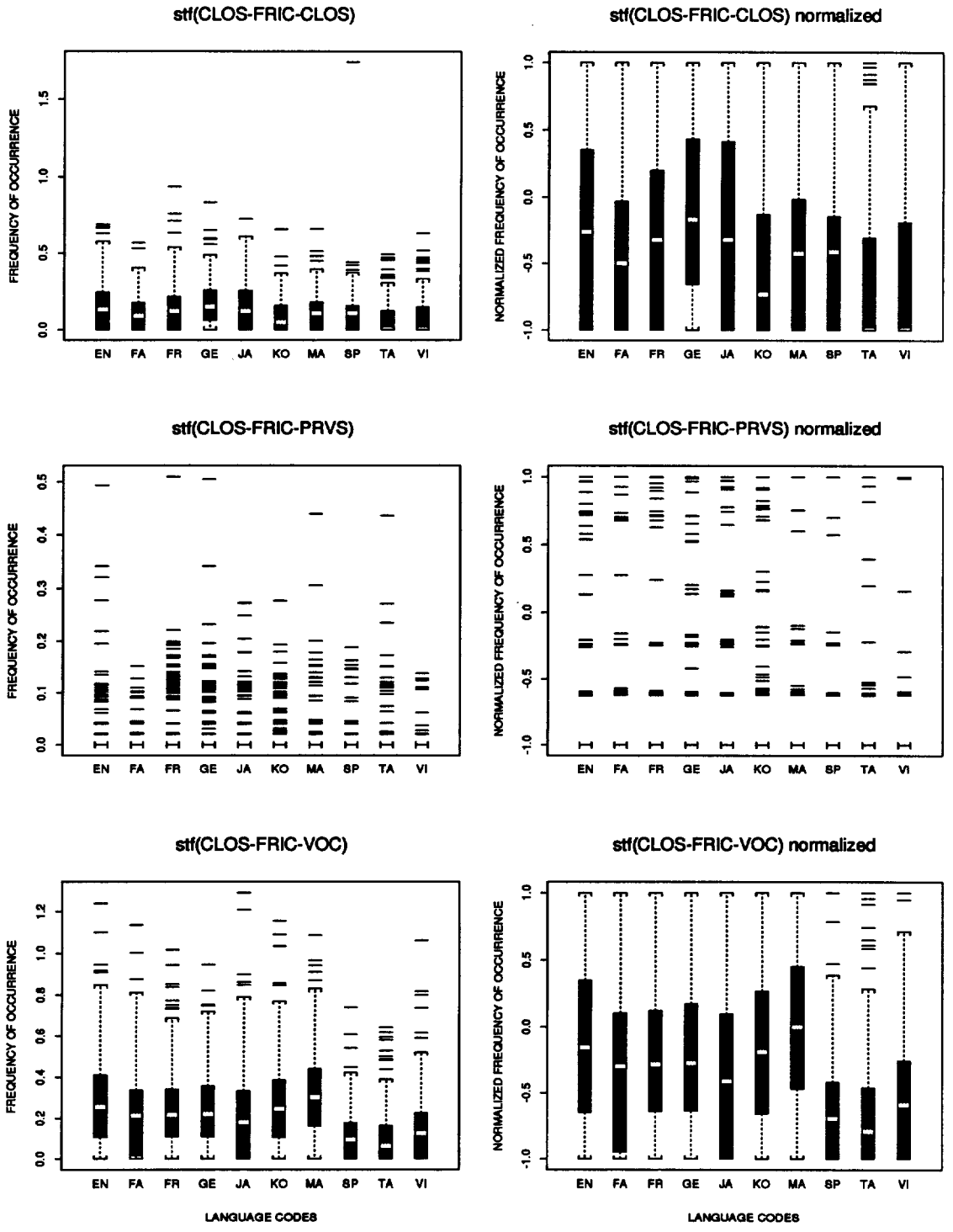

Figure C.1: Boxplots of Segment-triple Frequency (STF) Features

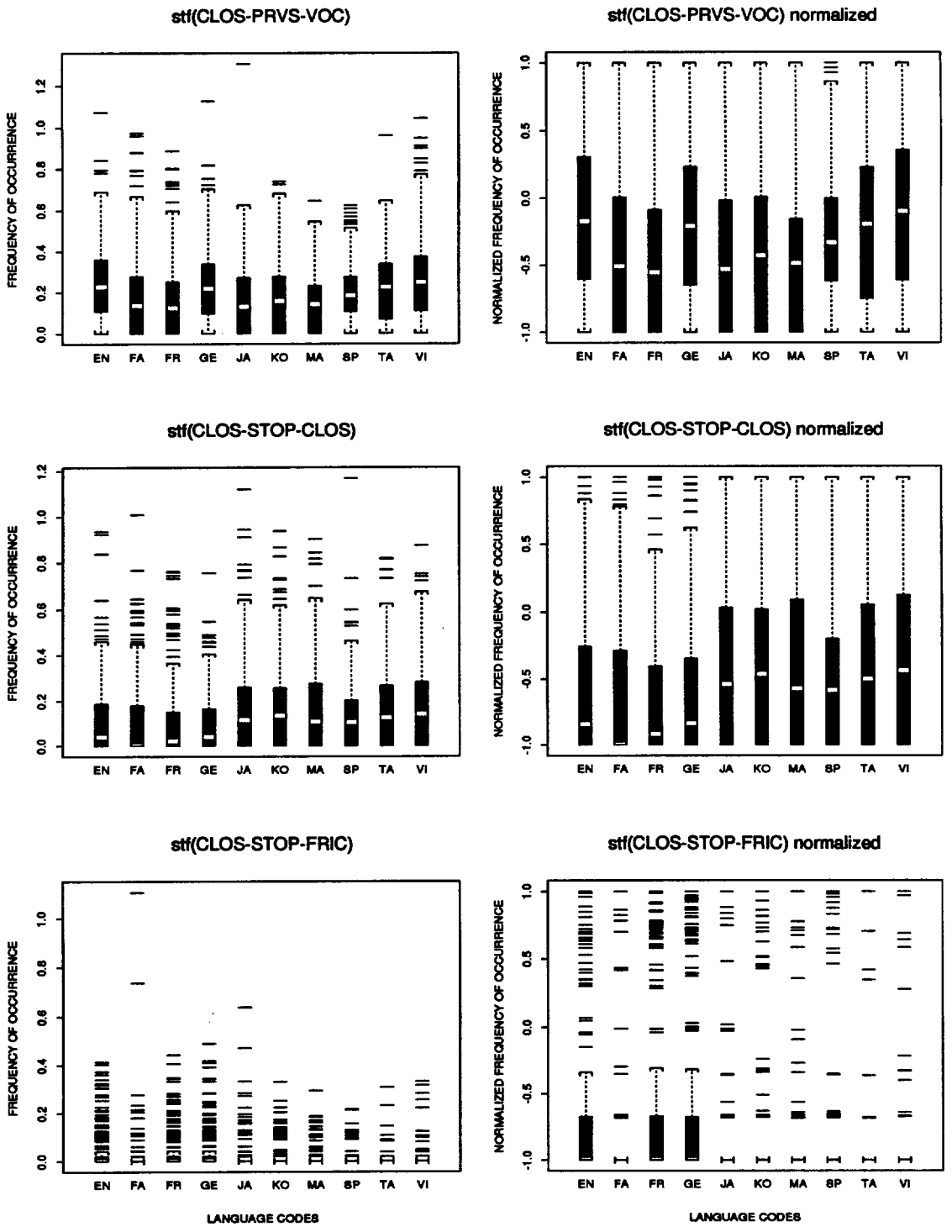

Figure C.2: Boxplots of STF Features (continued)

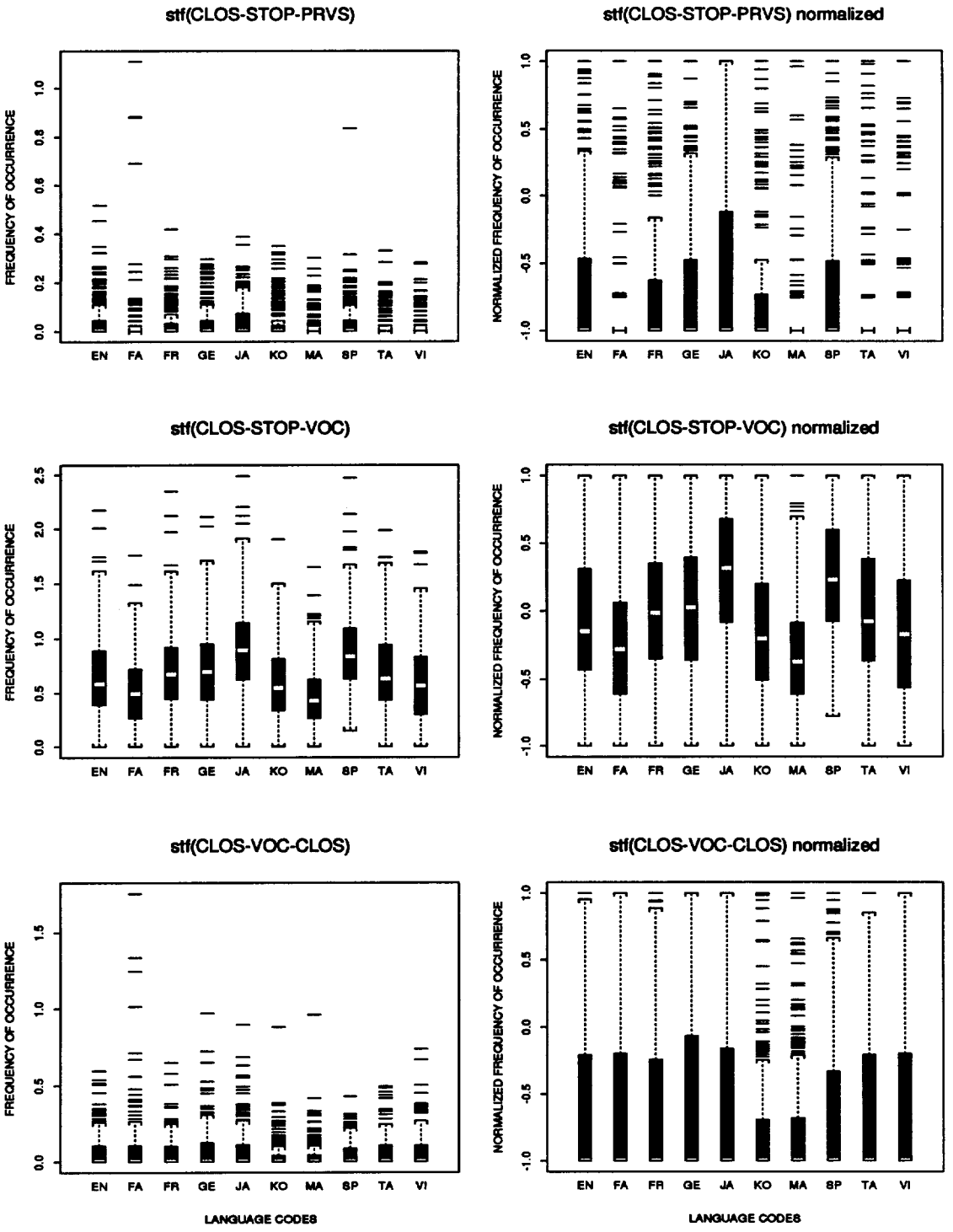

Figure C.3: Boxplots of STF Features (continued)

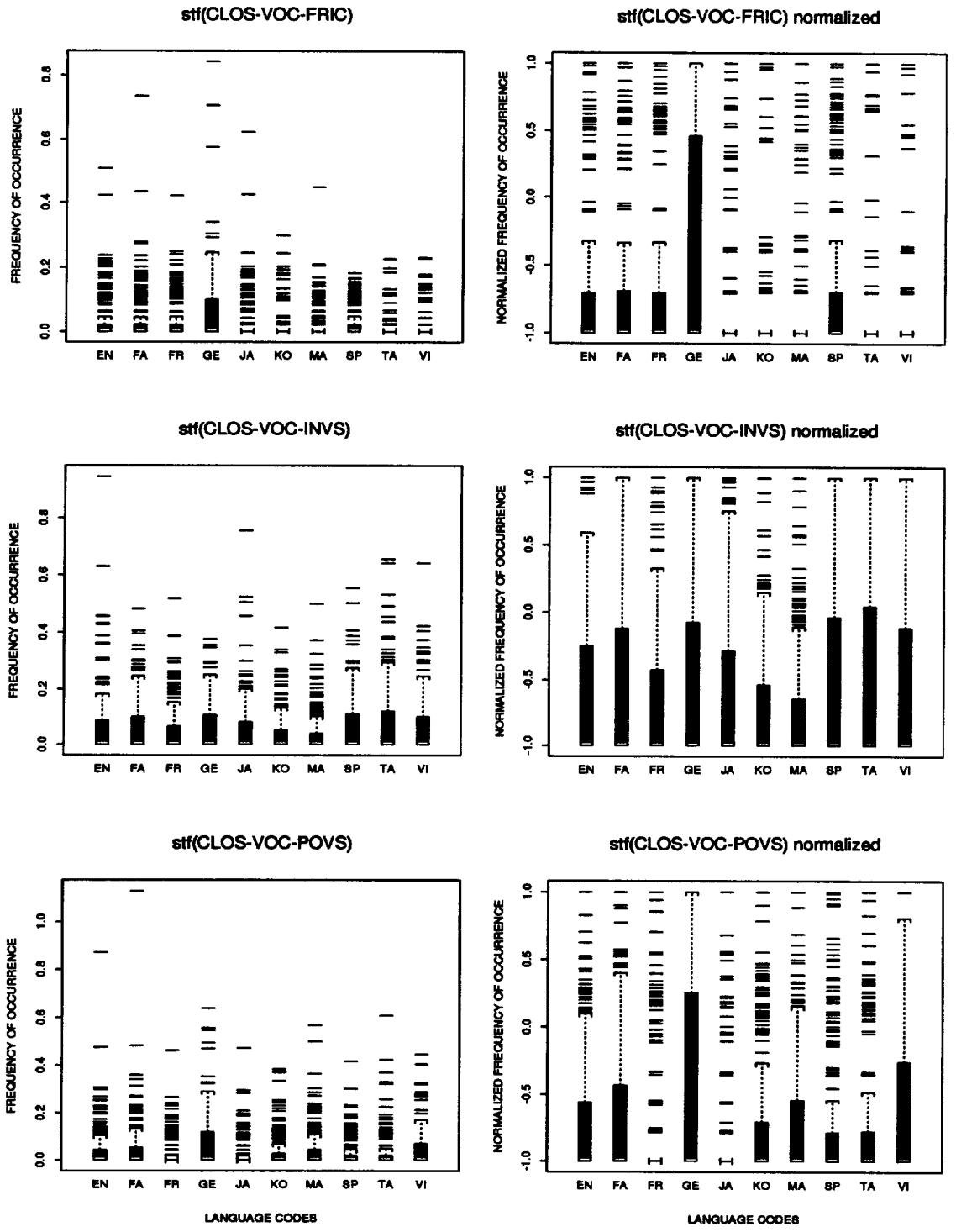

Figure C.4: Boxplots of STF Features (continued)

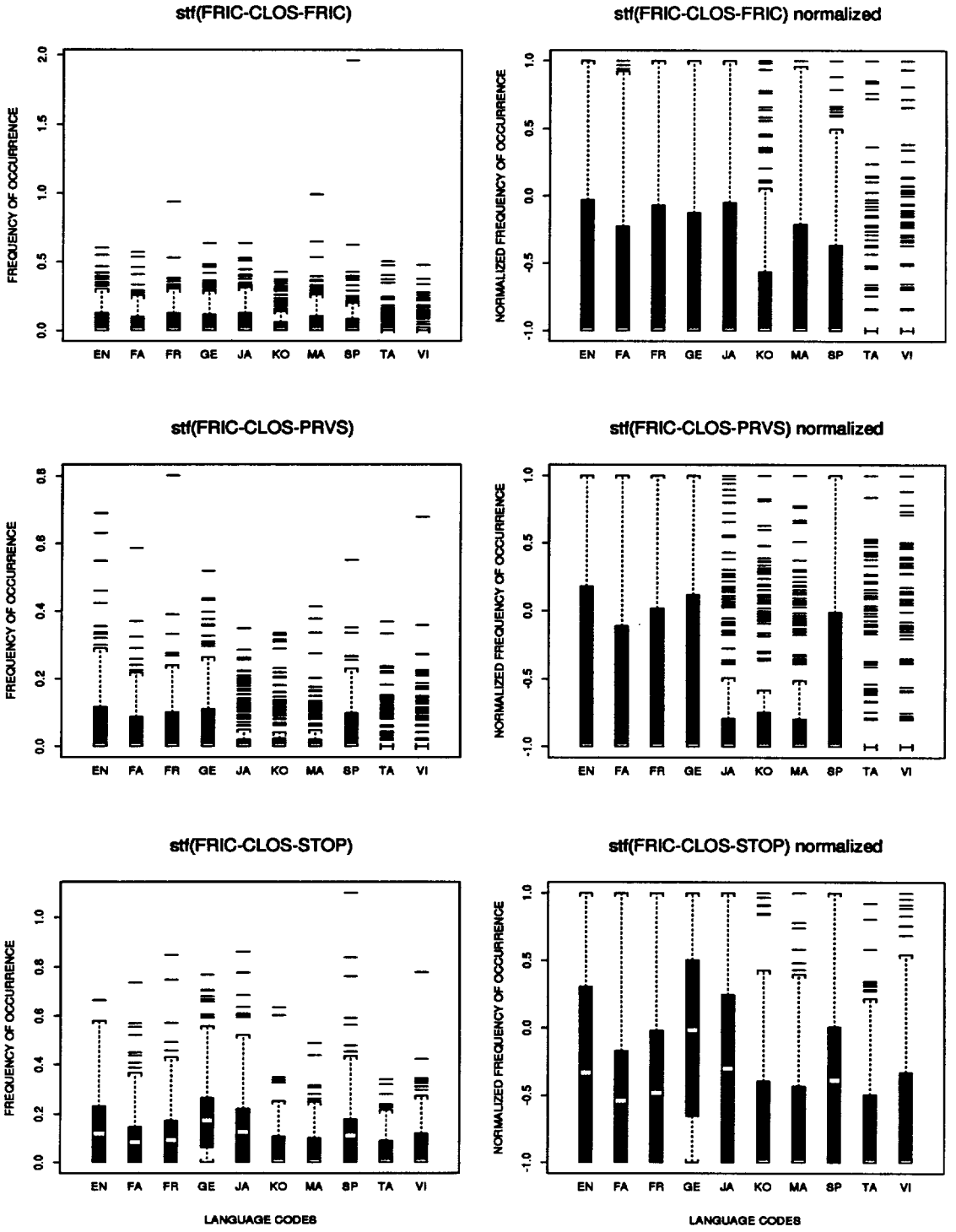

Figure C.5: Boxplots of STF Features (continued)

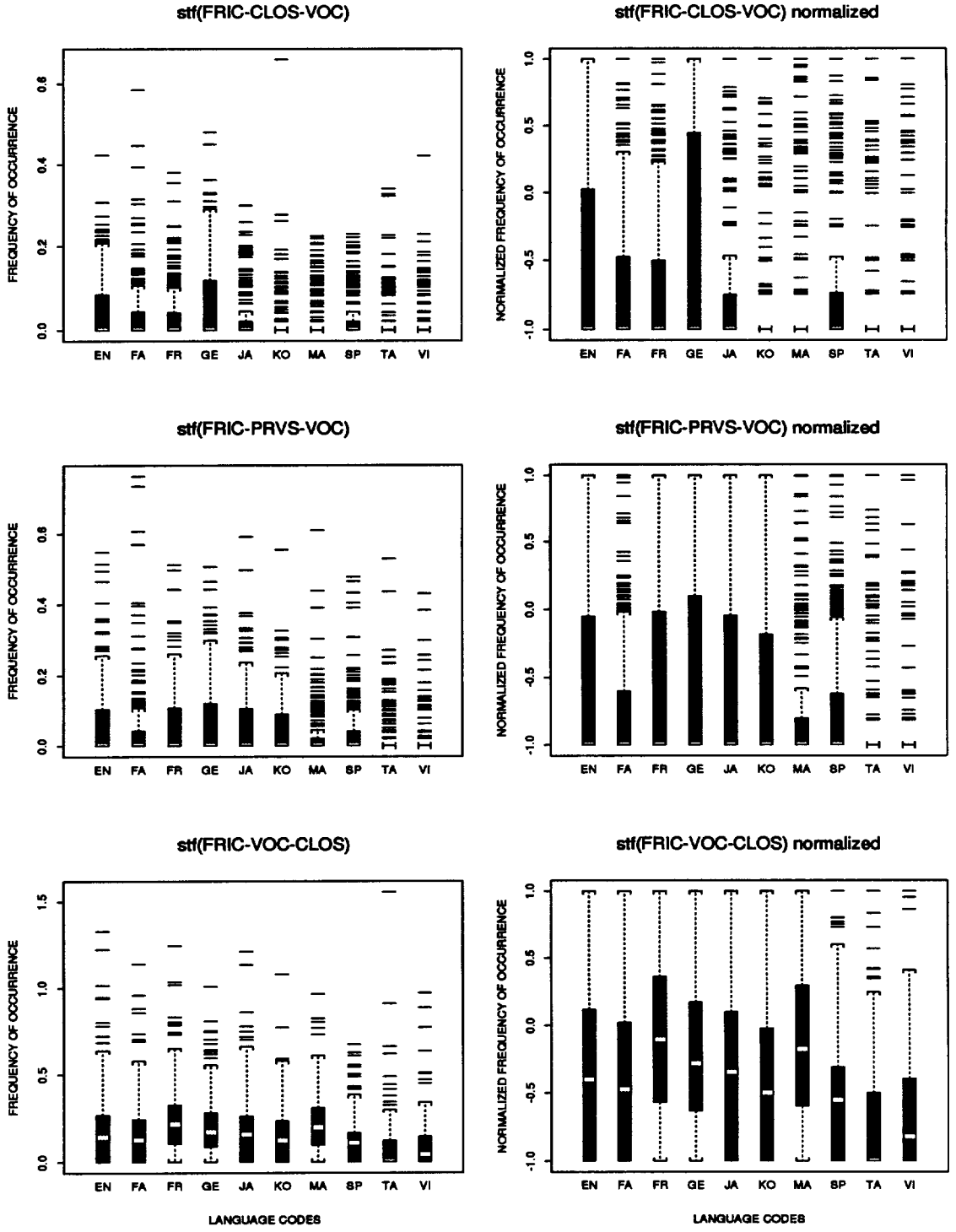

Figure C.6: Boxplots of STF Features (continued)

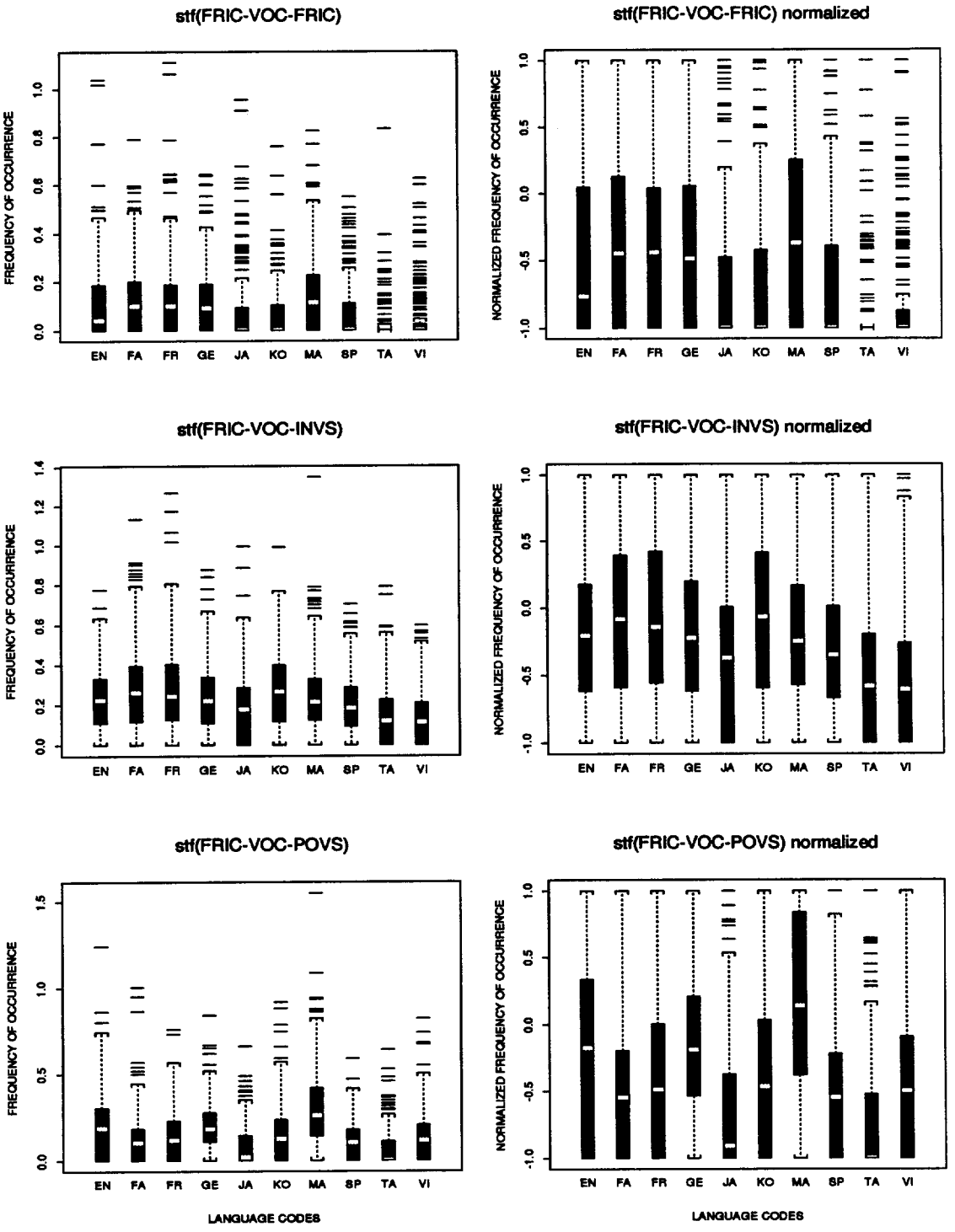

Figure C.7: Boxplots of STF Features (continued)

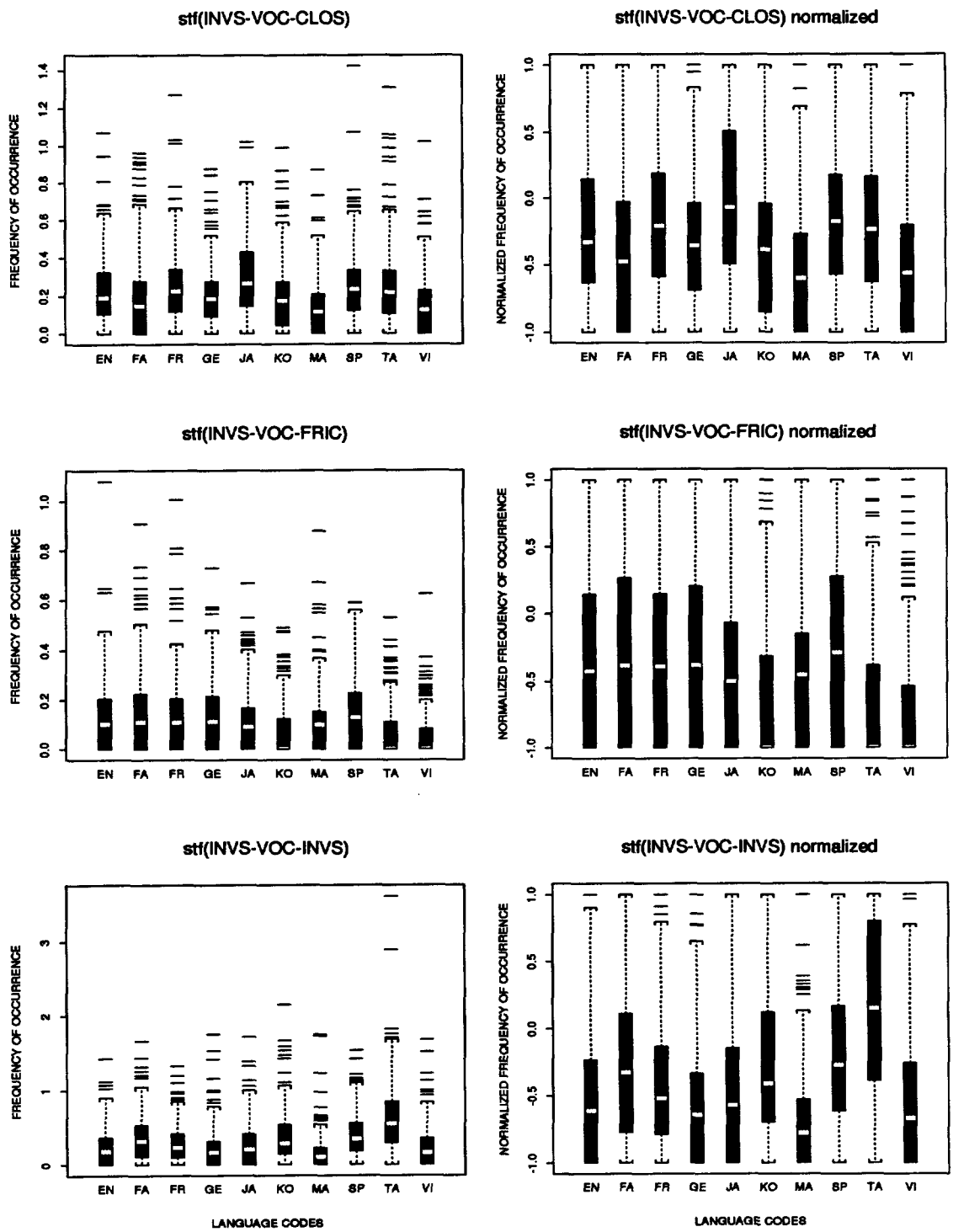

Figure C.8: Boxplots of STF Features (continued)

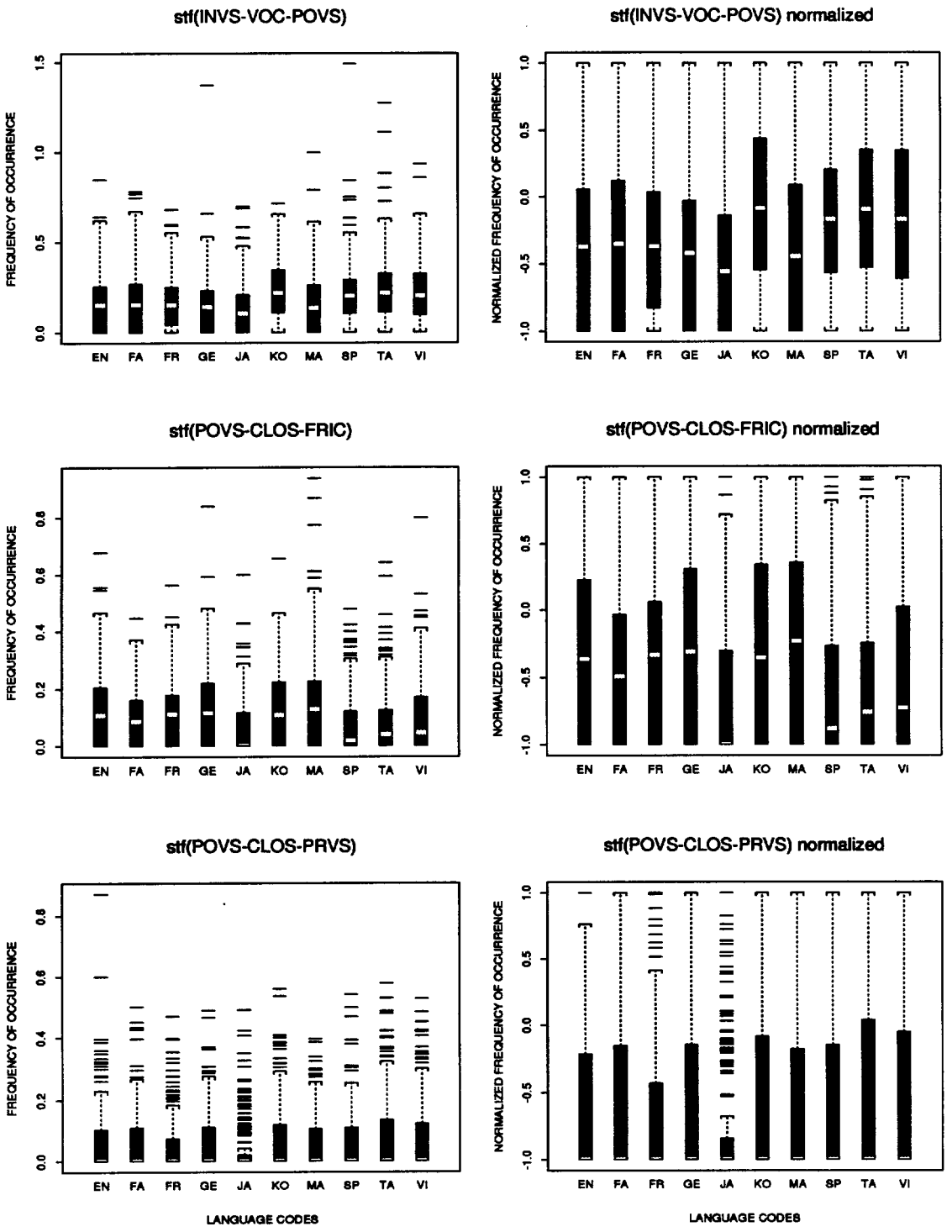

Figure C.9: Boxplots of STF Features (continued)

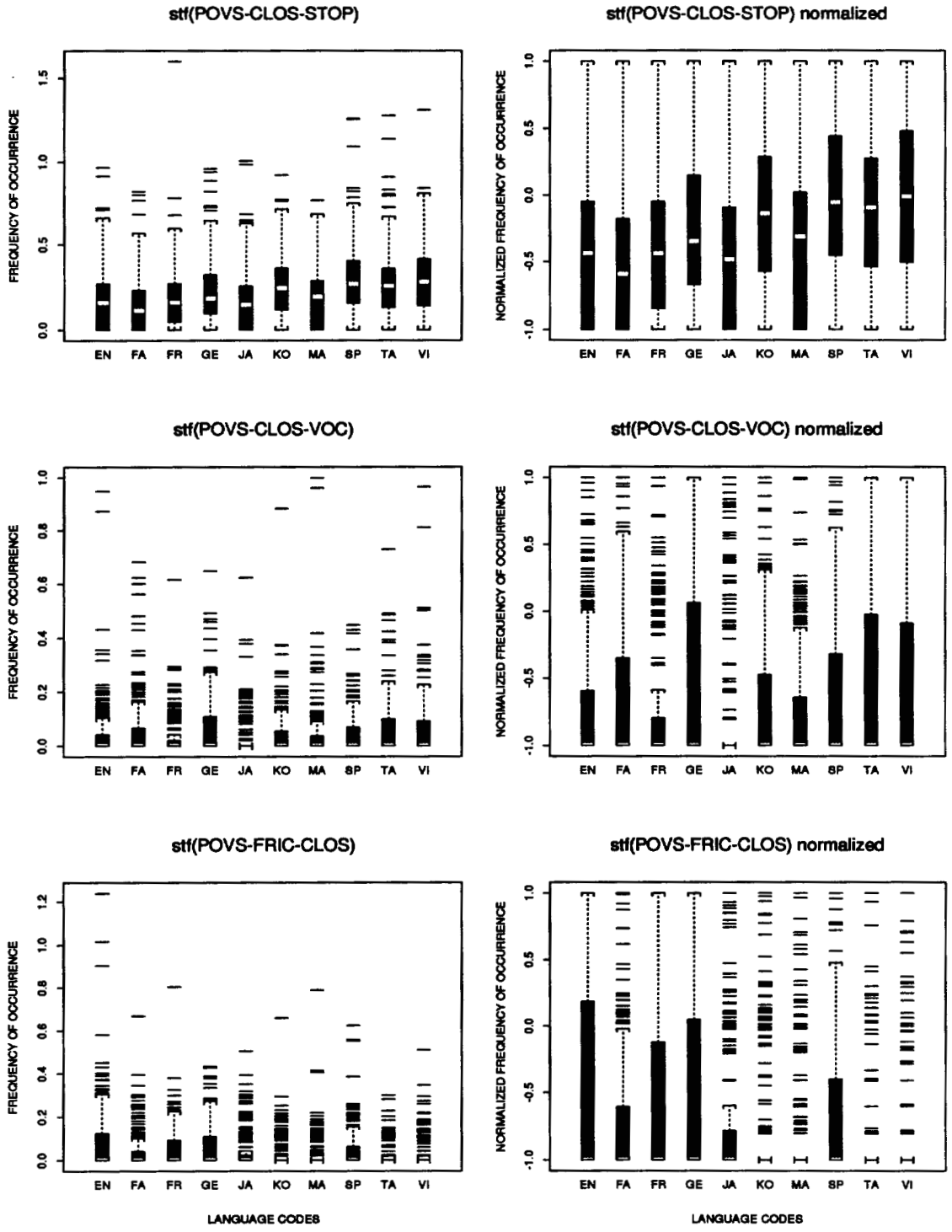

Figure C.10: Boxplots of STF Features (continued)

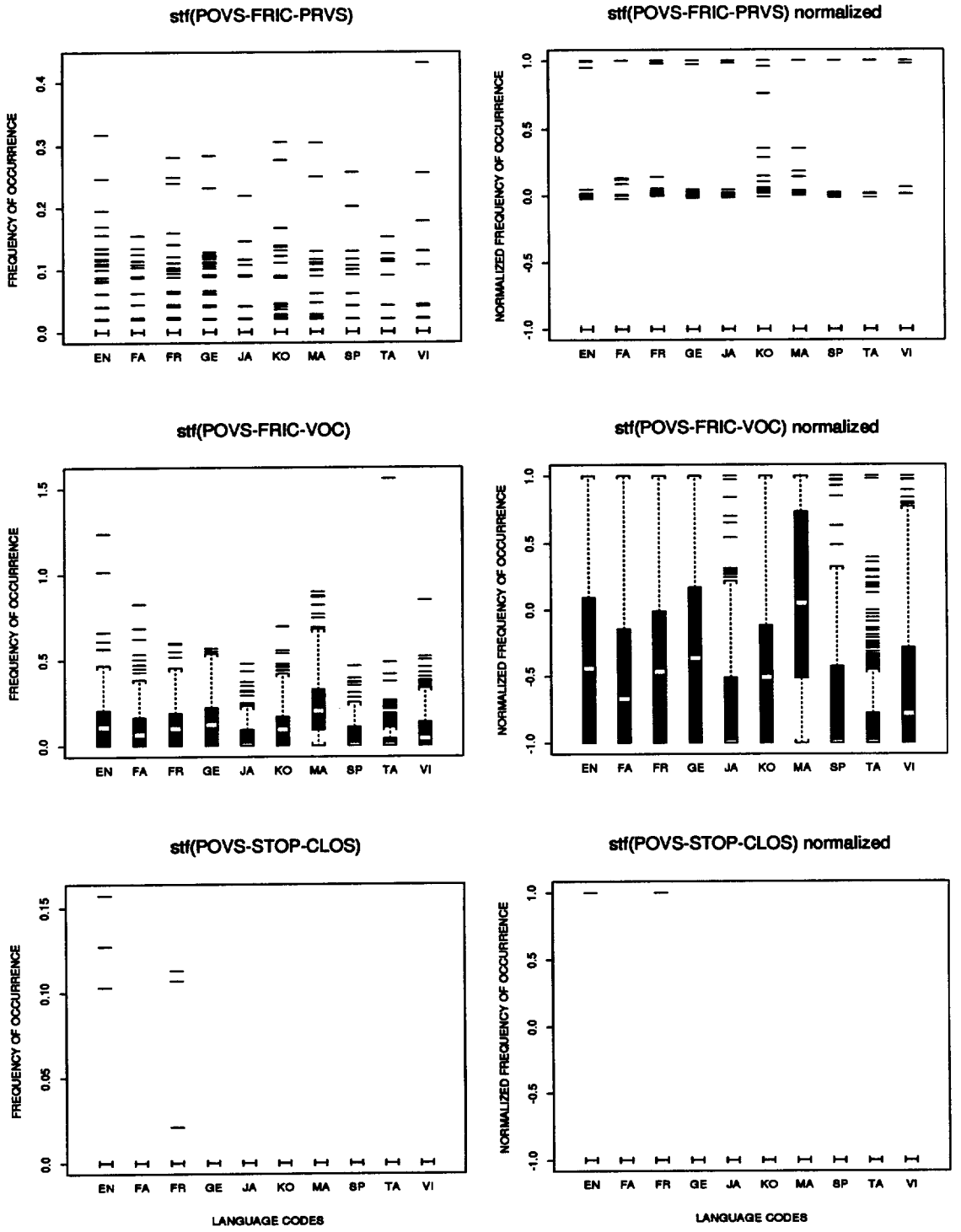

Figure C.11: Boxplots of STF Features (continued)

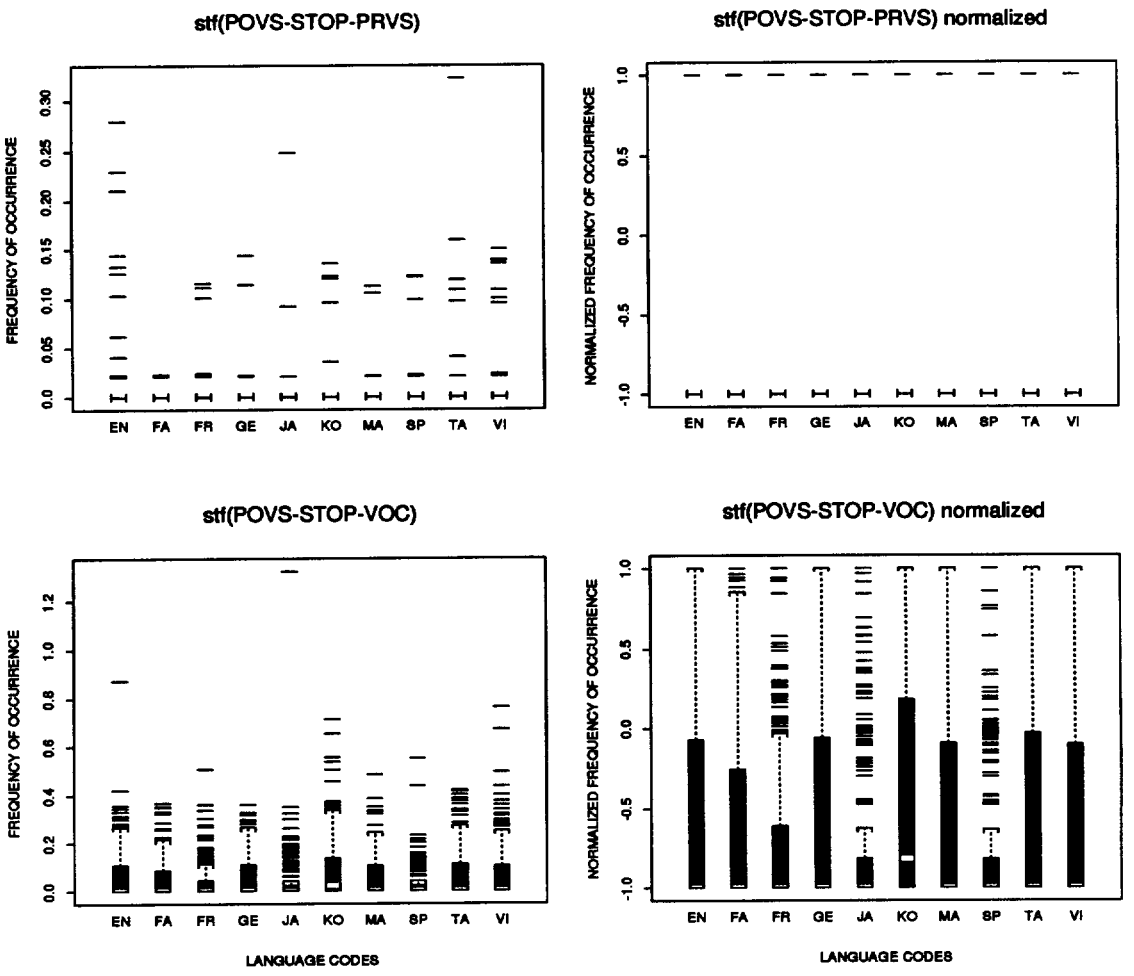

Figure C.12: Boxplots of STF Features (continued)

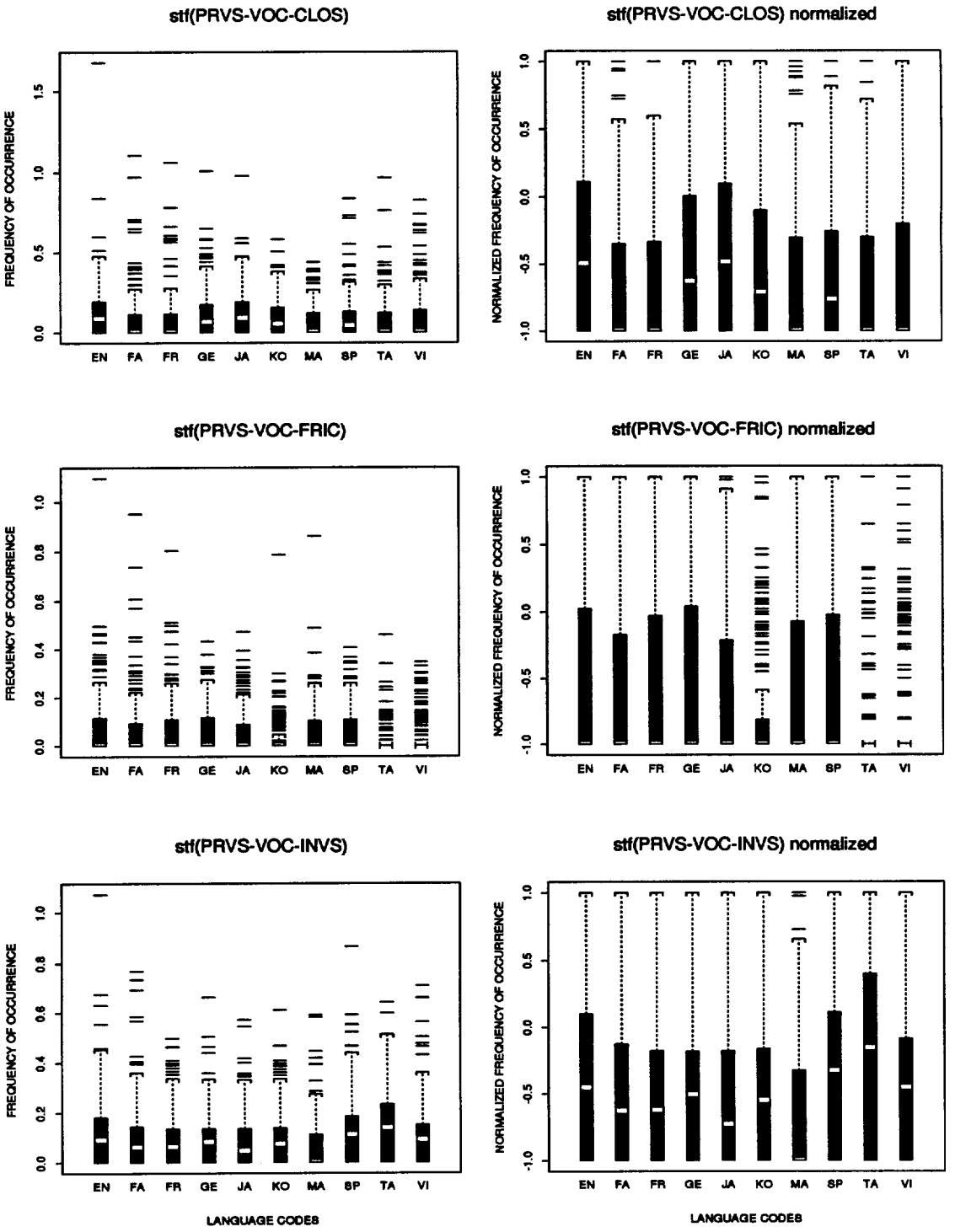

Figure C.13: Boxplots of STF Features (continued)

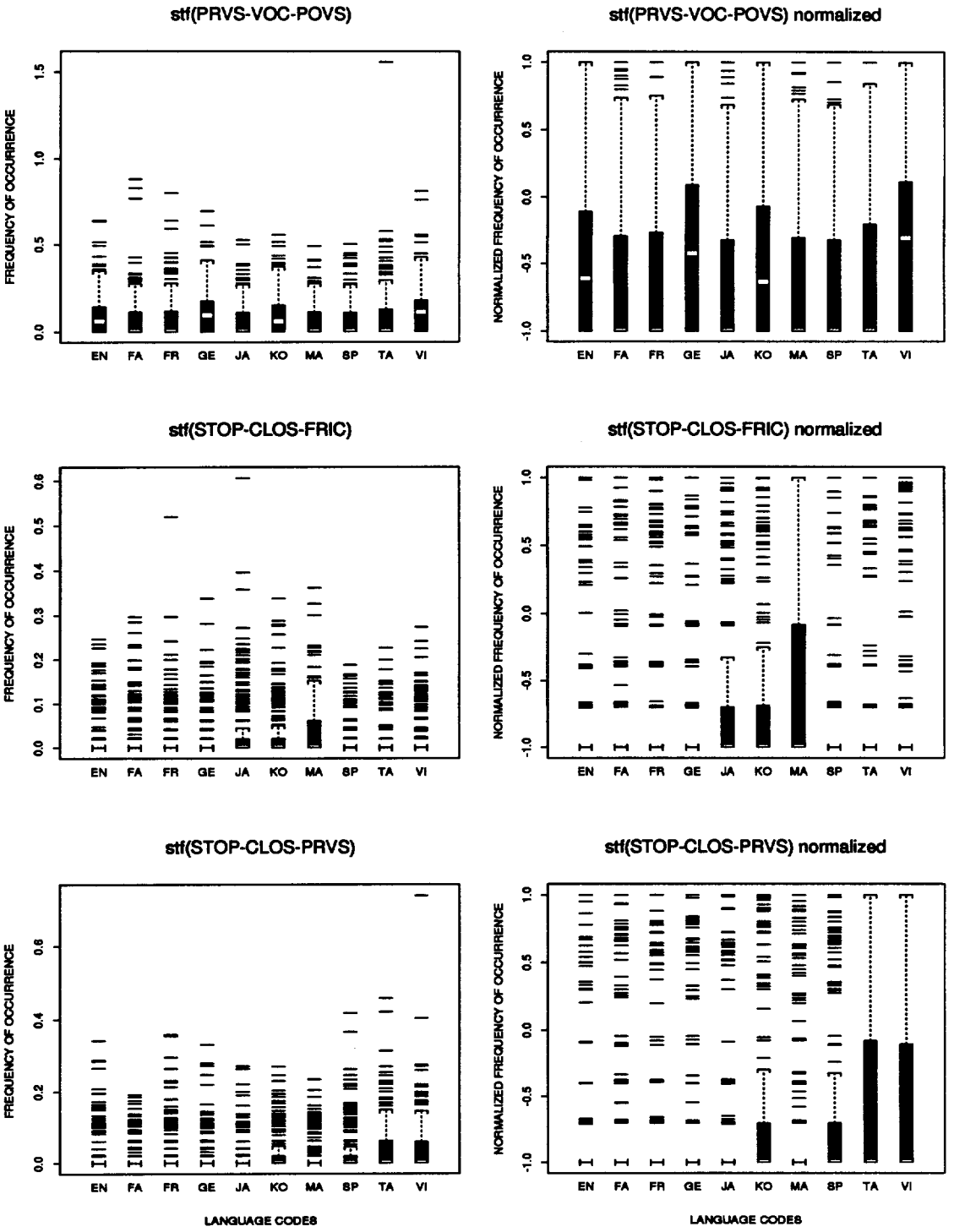

Figure C.14: Boxplots of STF Features (continued)

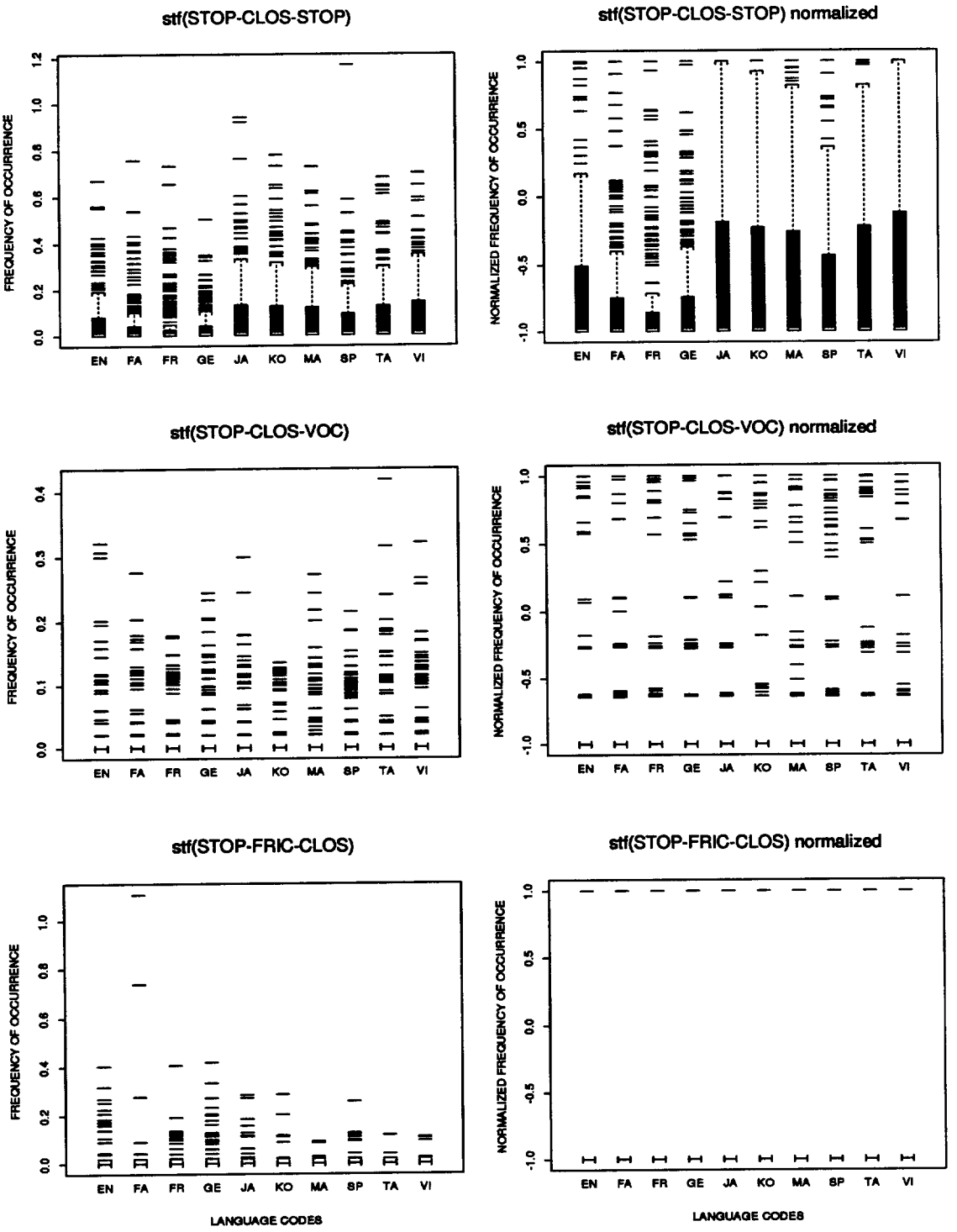

Figure C.15: Boxplots of STF Features (continued)

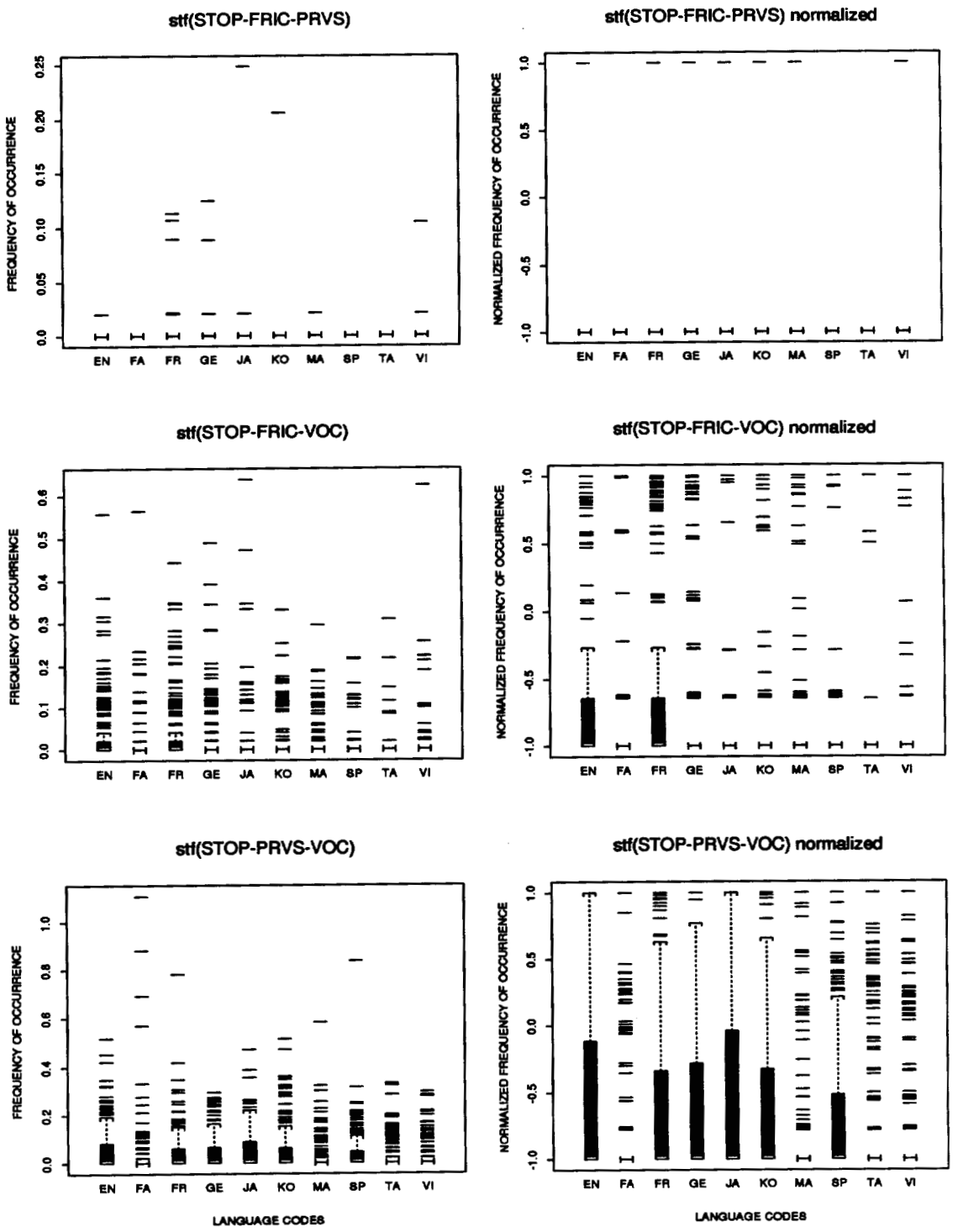

Figure C.16: Boxplots of STF Features (continued)

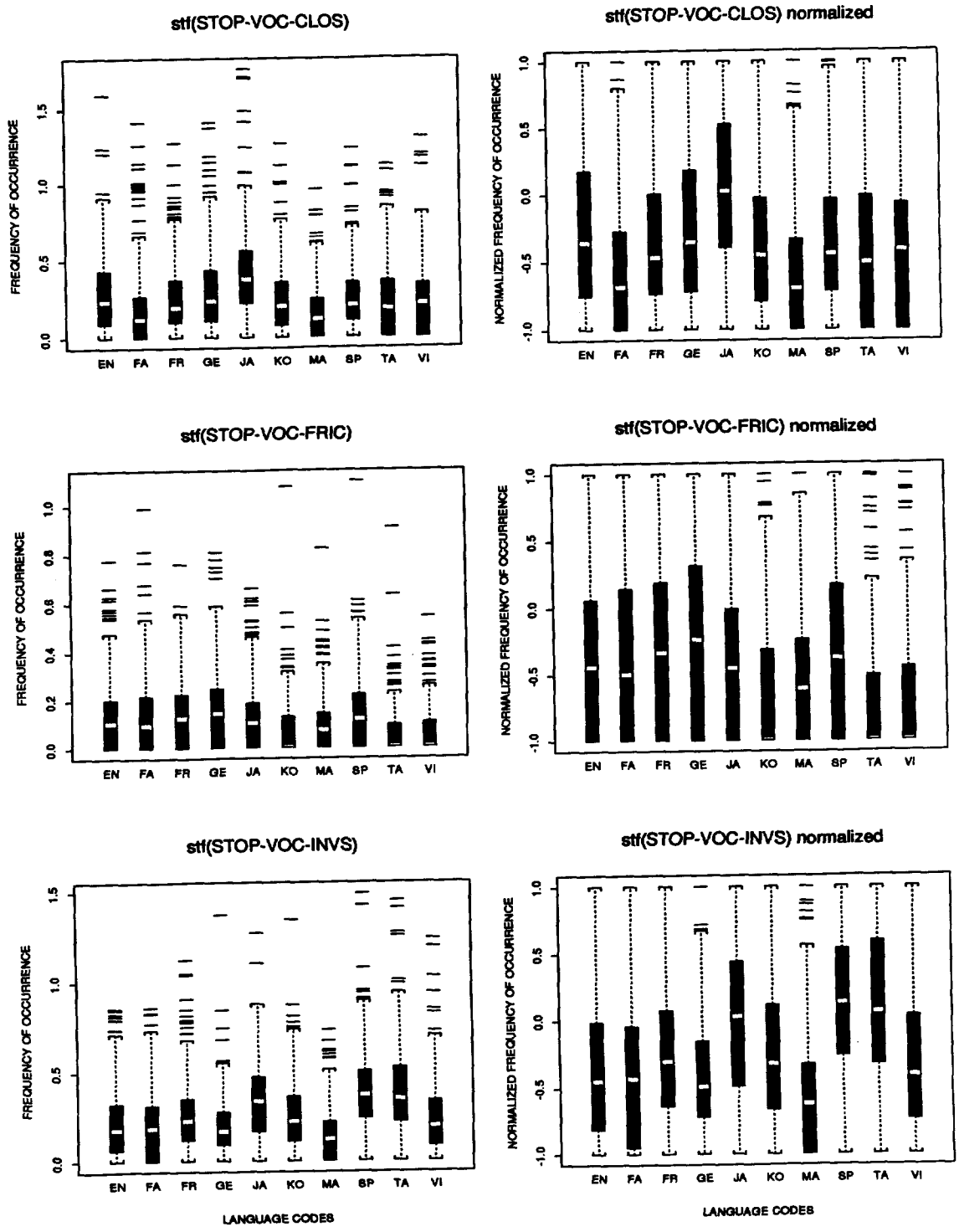

Figure C.17: Boxplots of STF Features (continued)

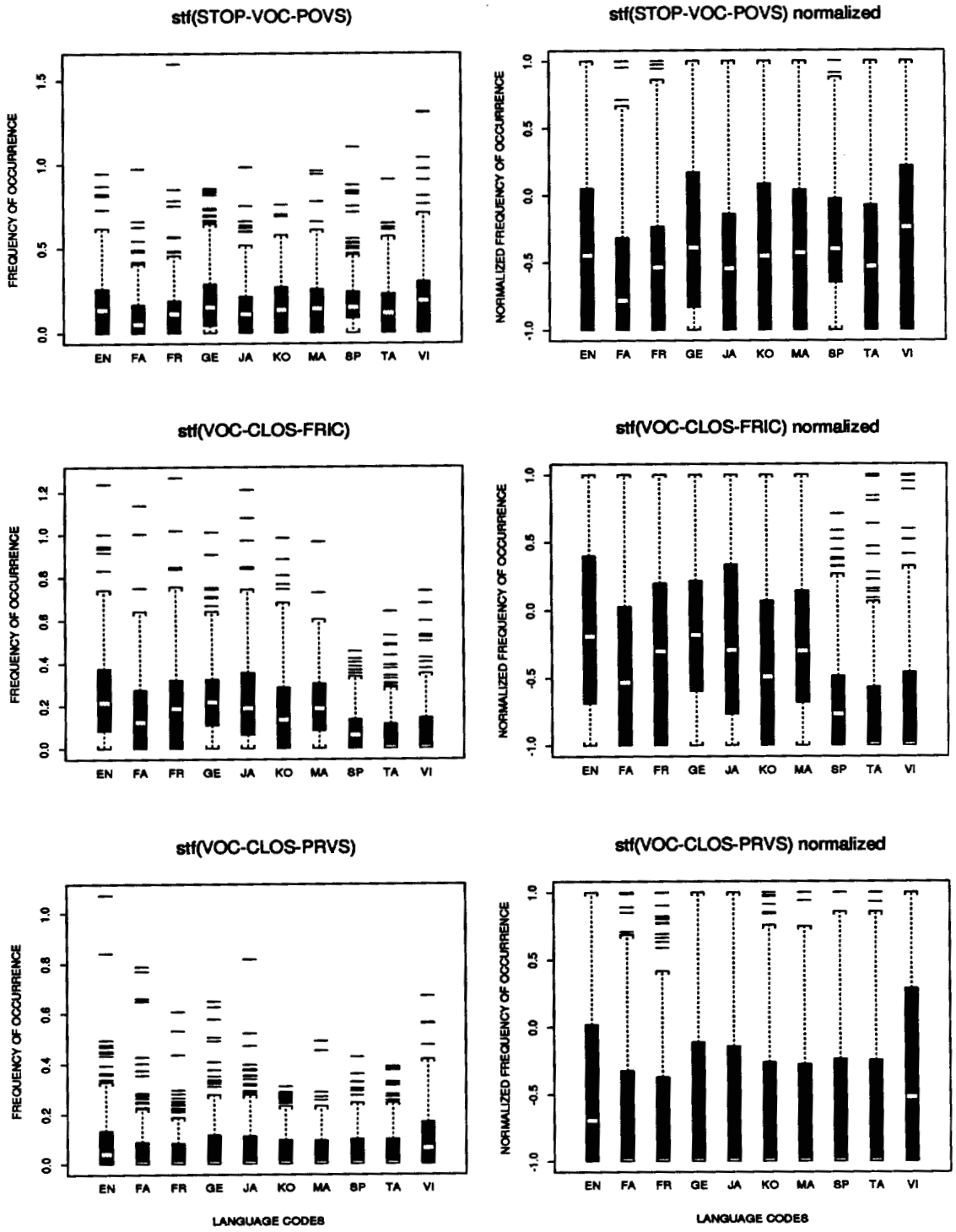

Figure C.18: Boxplots of STF Features (continued)

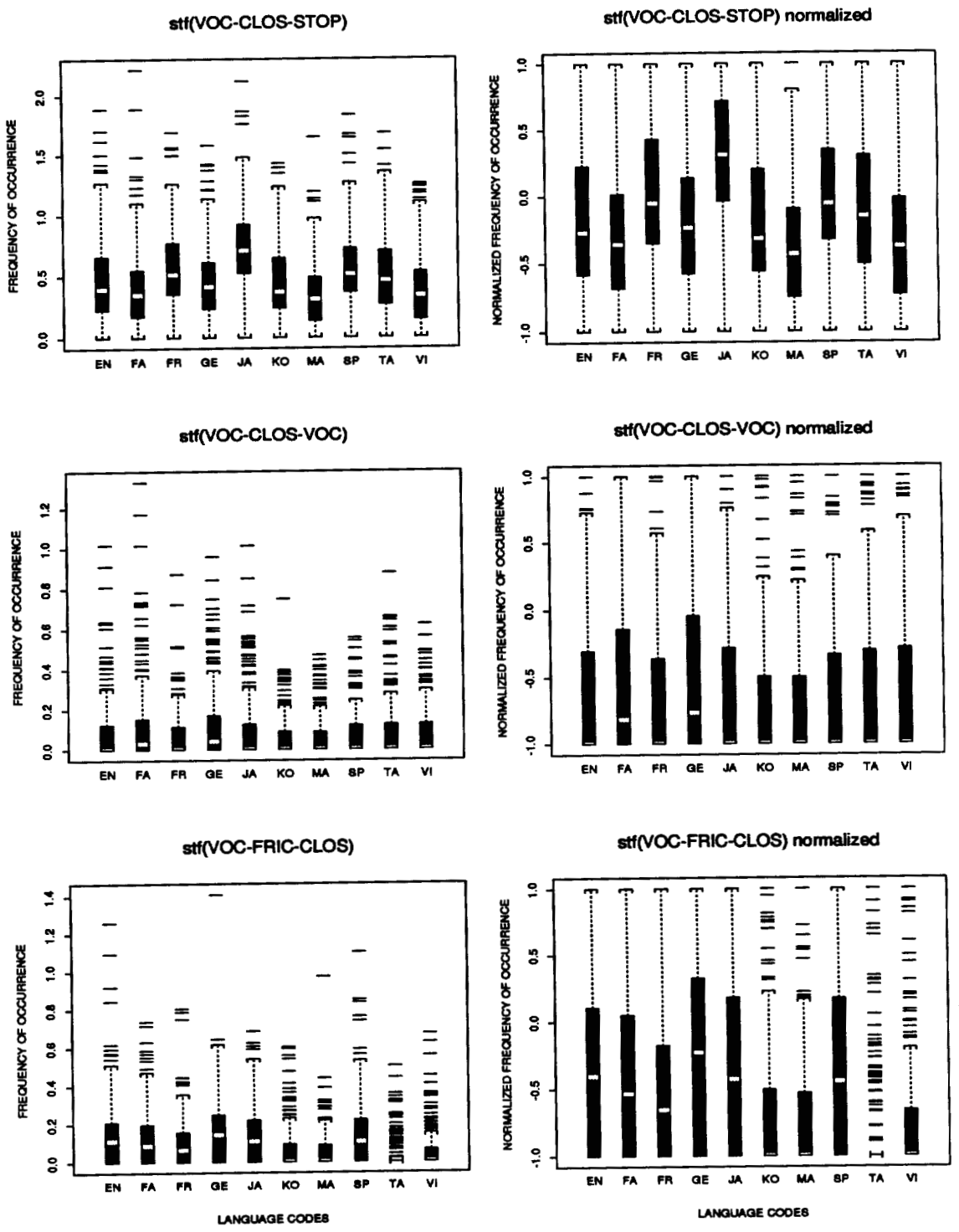

Figure C.19: Boxplots of STF Features (continued)

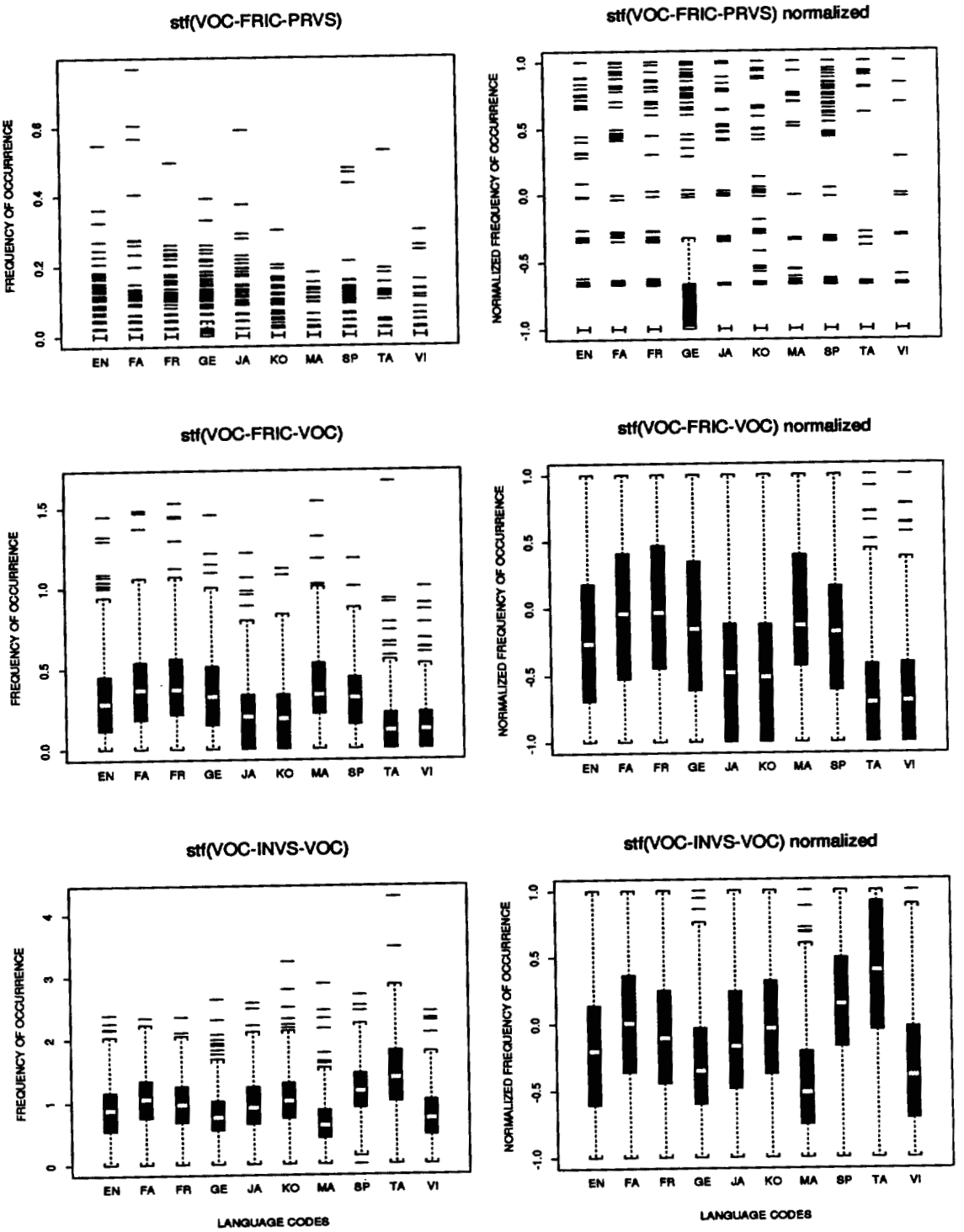

Figure C.20: Boxplots of STF Features (continued)

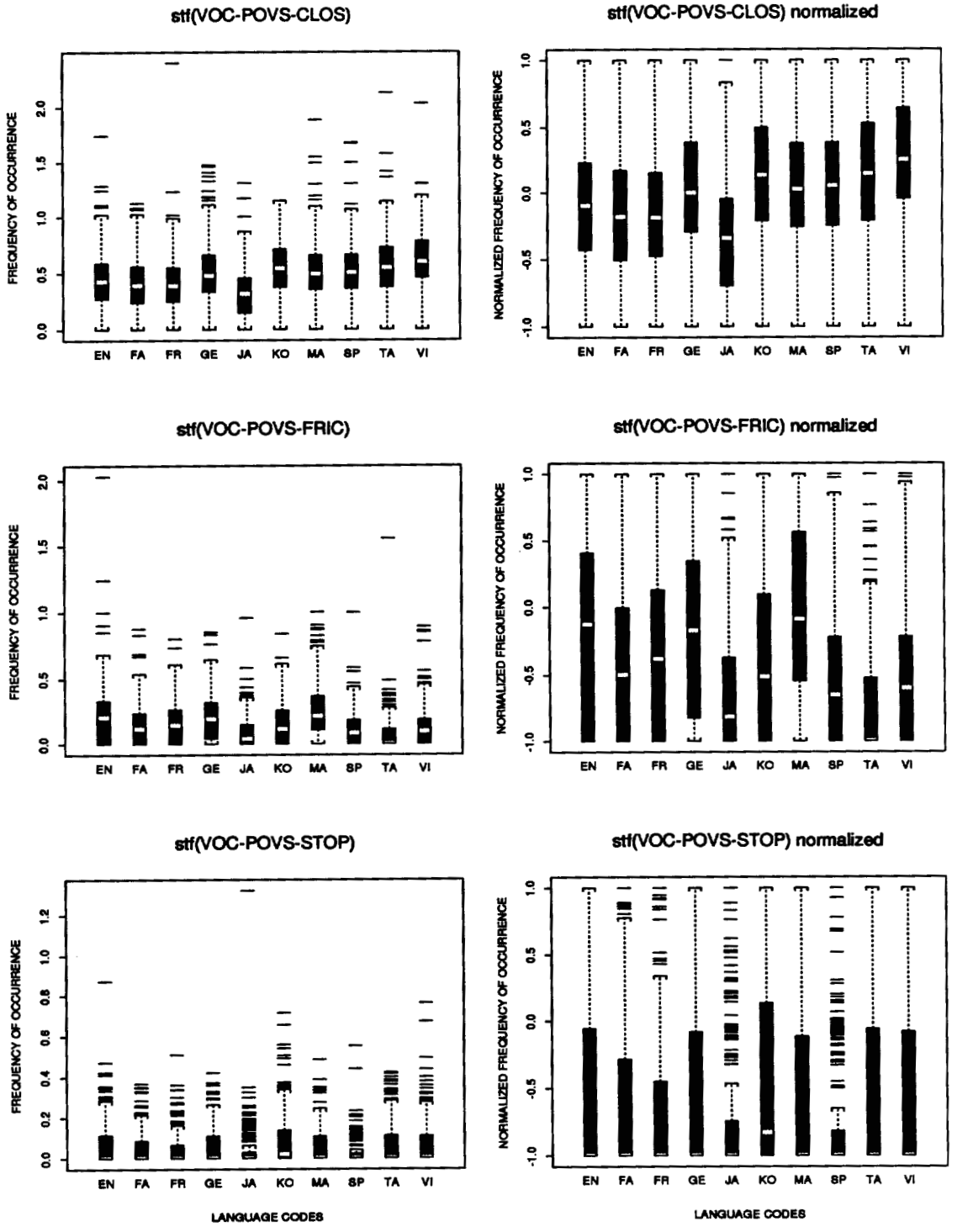

Figure C.21: Boxplots of STF Features (continued)

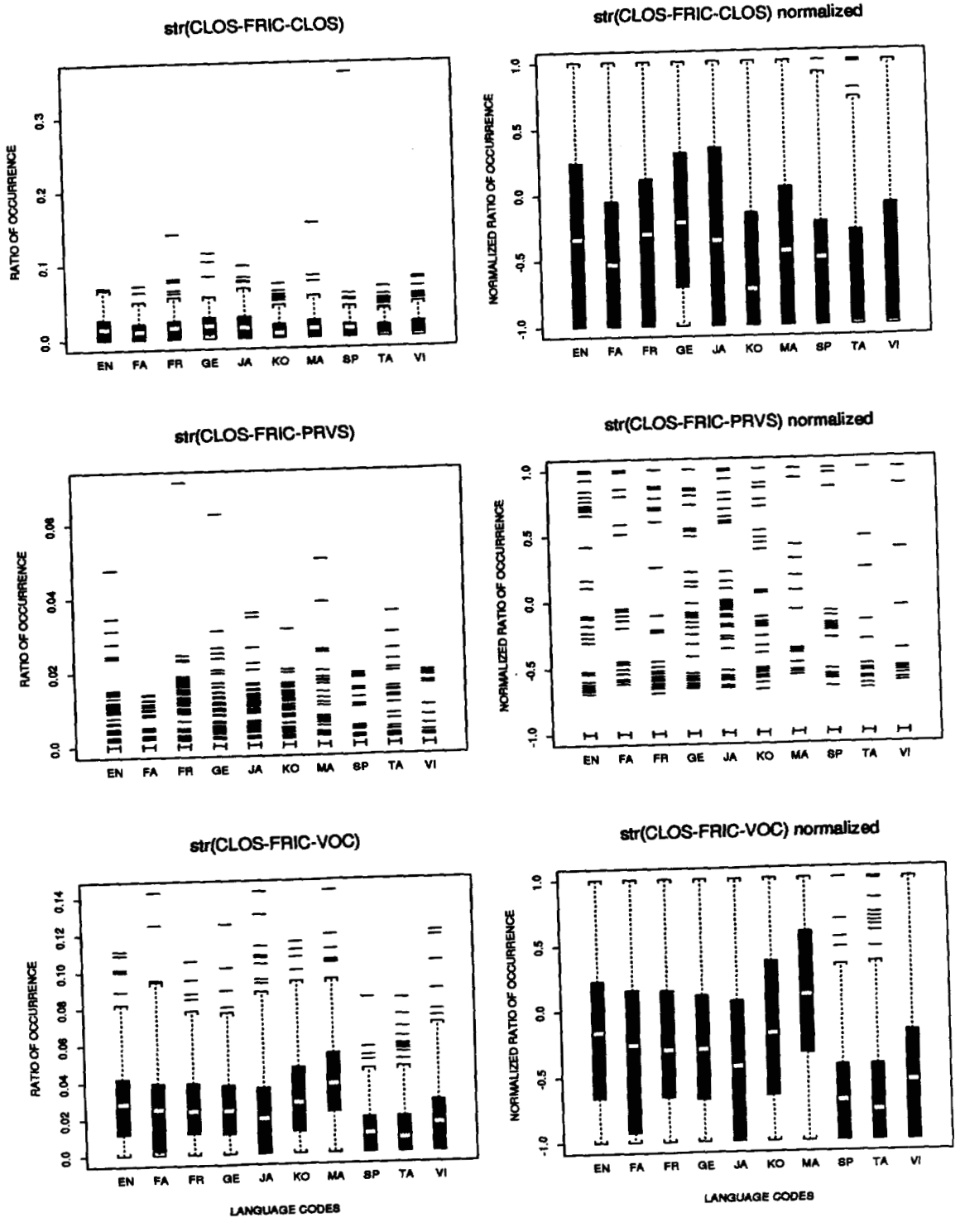

Figure C.22: Boxplots of Segment-triple Ratio (STR) Features

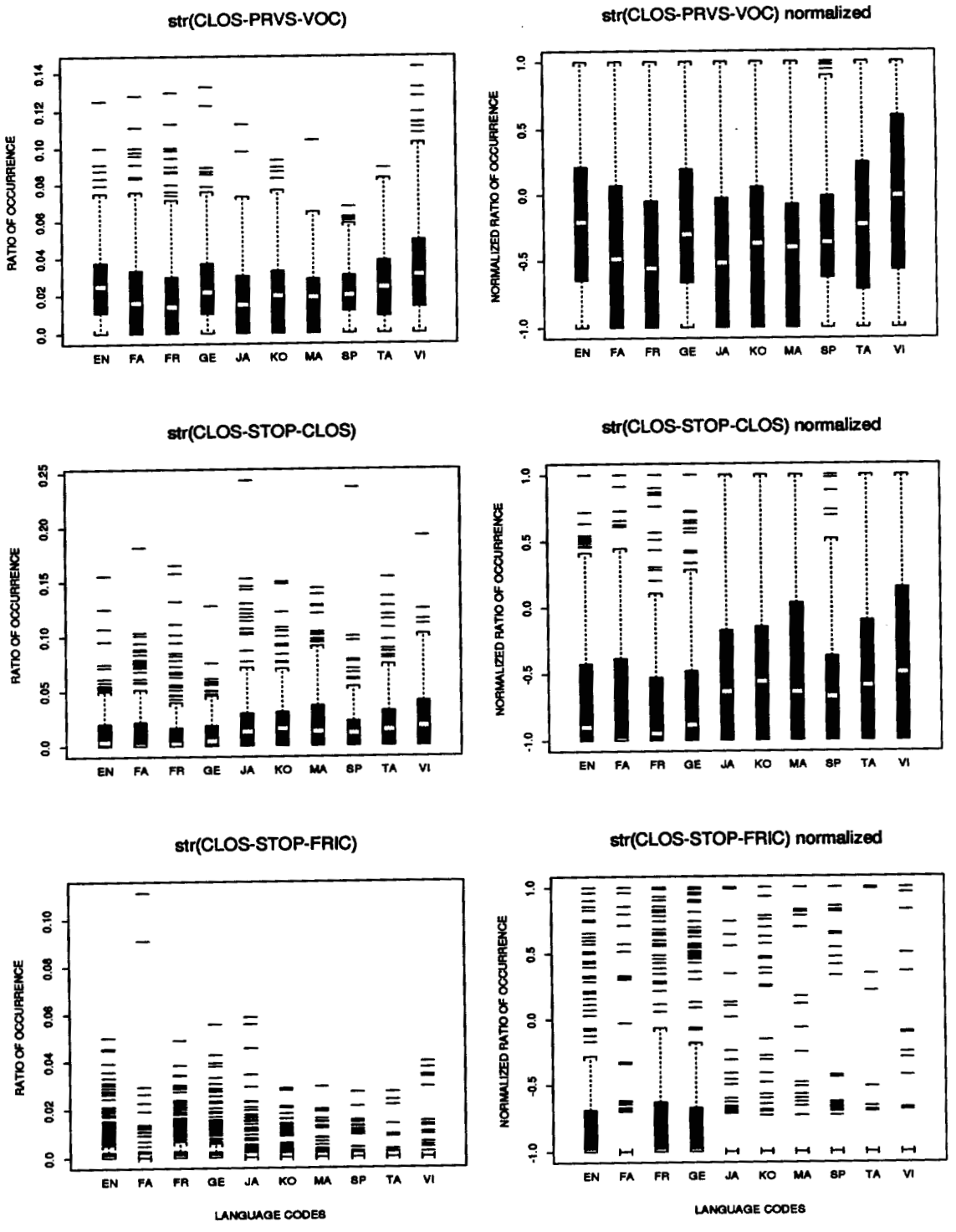

Figure C.23: Boxplots of STR Features (continued)

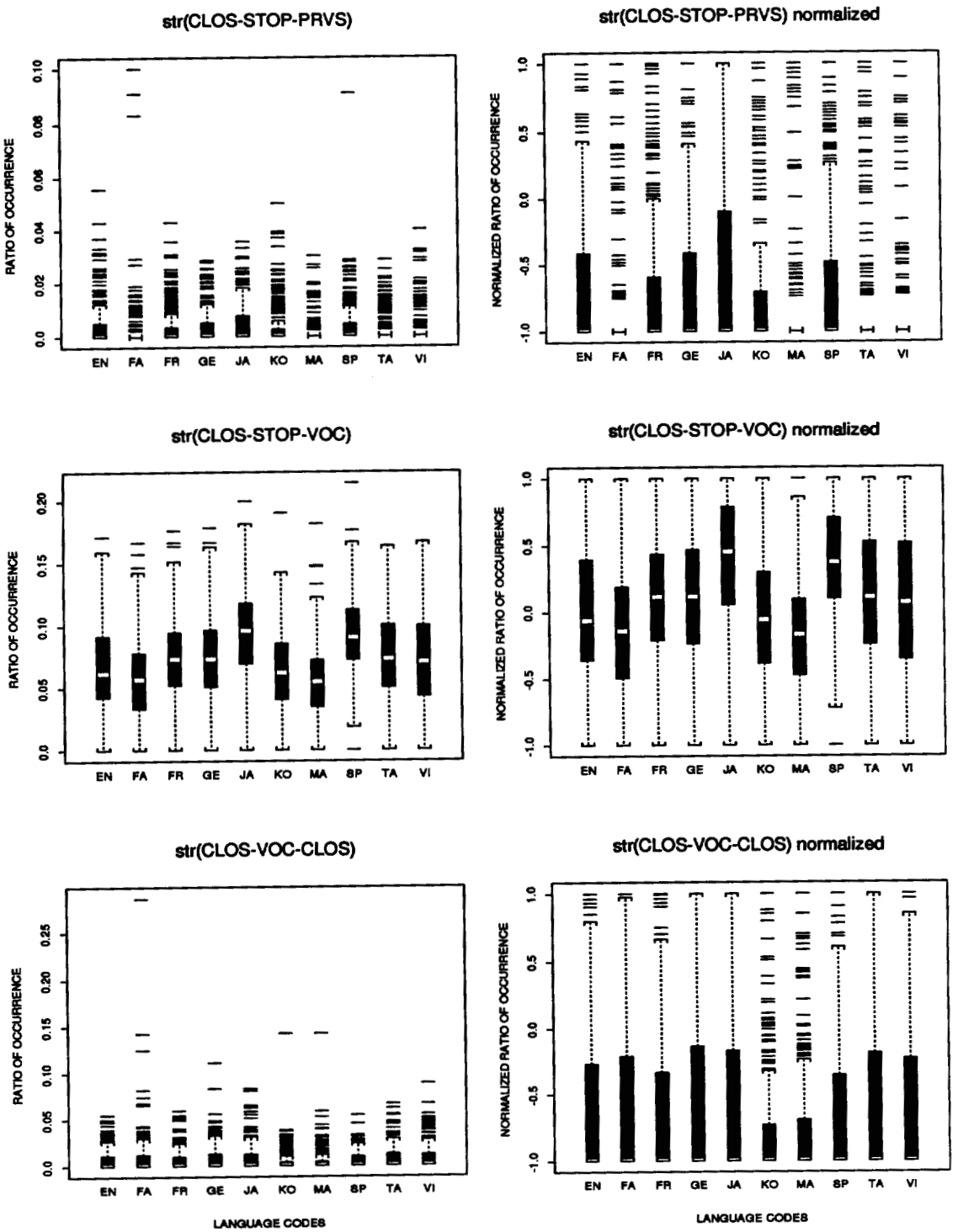

Figure C.24: Boxplots of STR Features (continued)

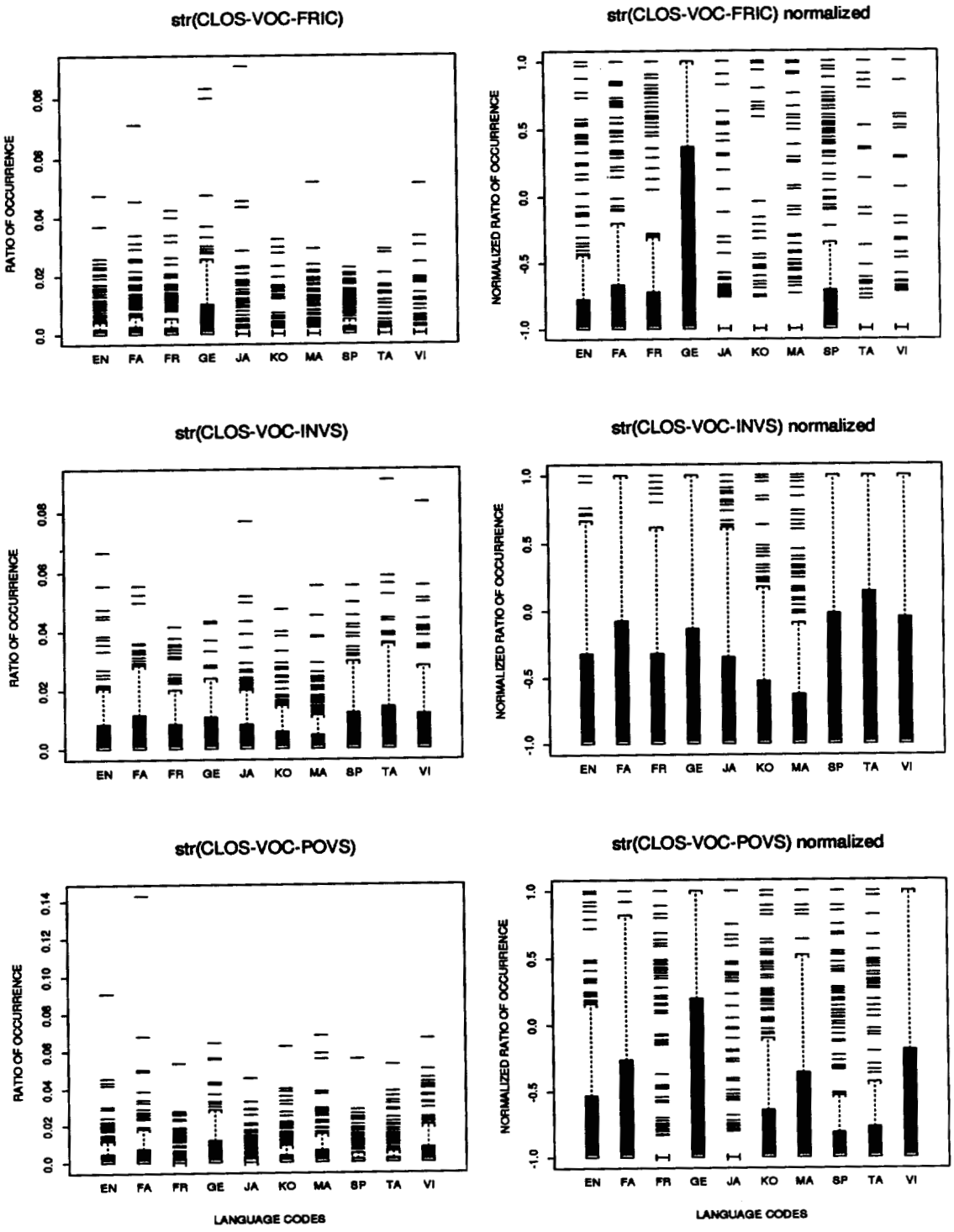

Figure C.25: Boxplots of STR Features (continued)

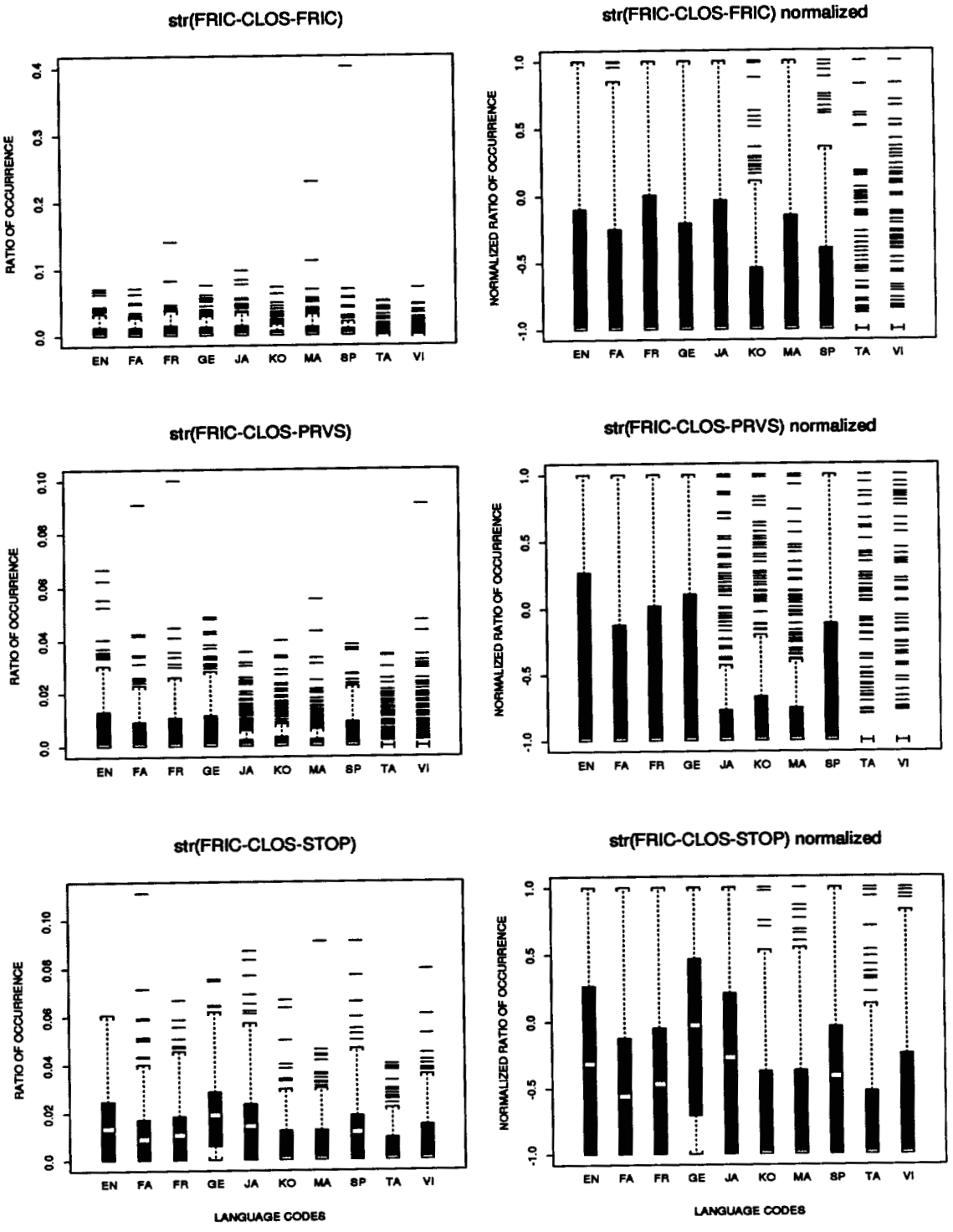

Figure C.26: Boxplots of STR Features (continued)

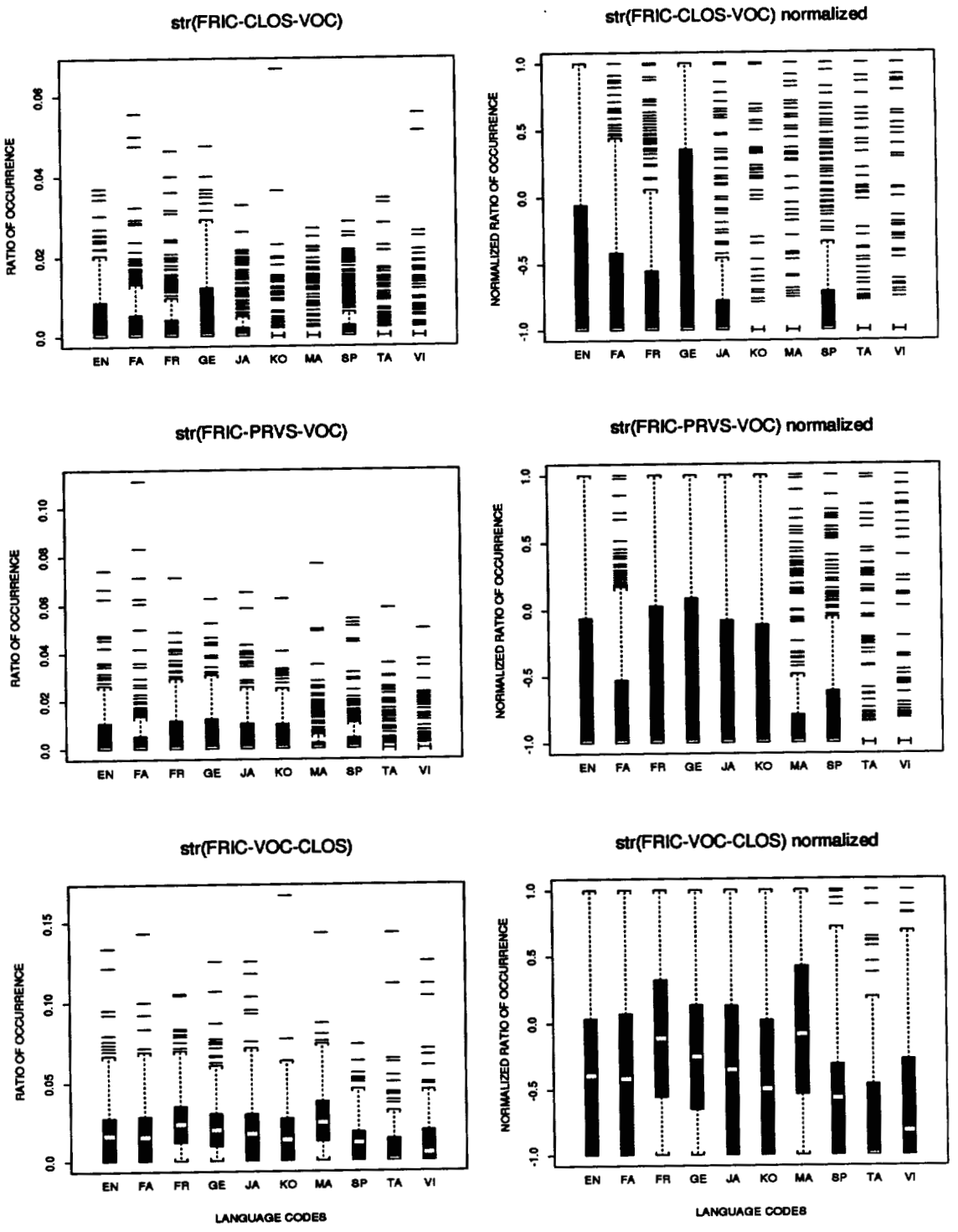

Figure C.27: Boxplots of STR Features (continued)

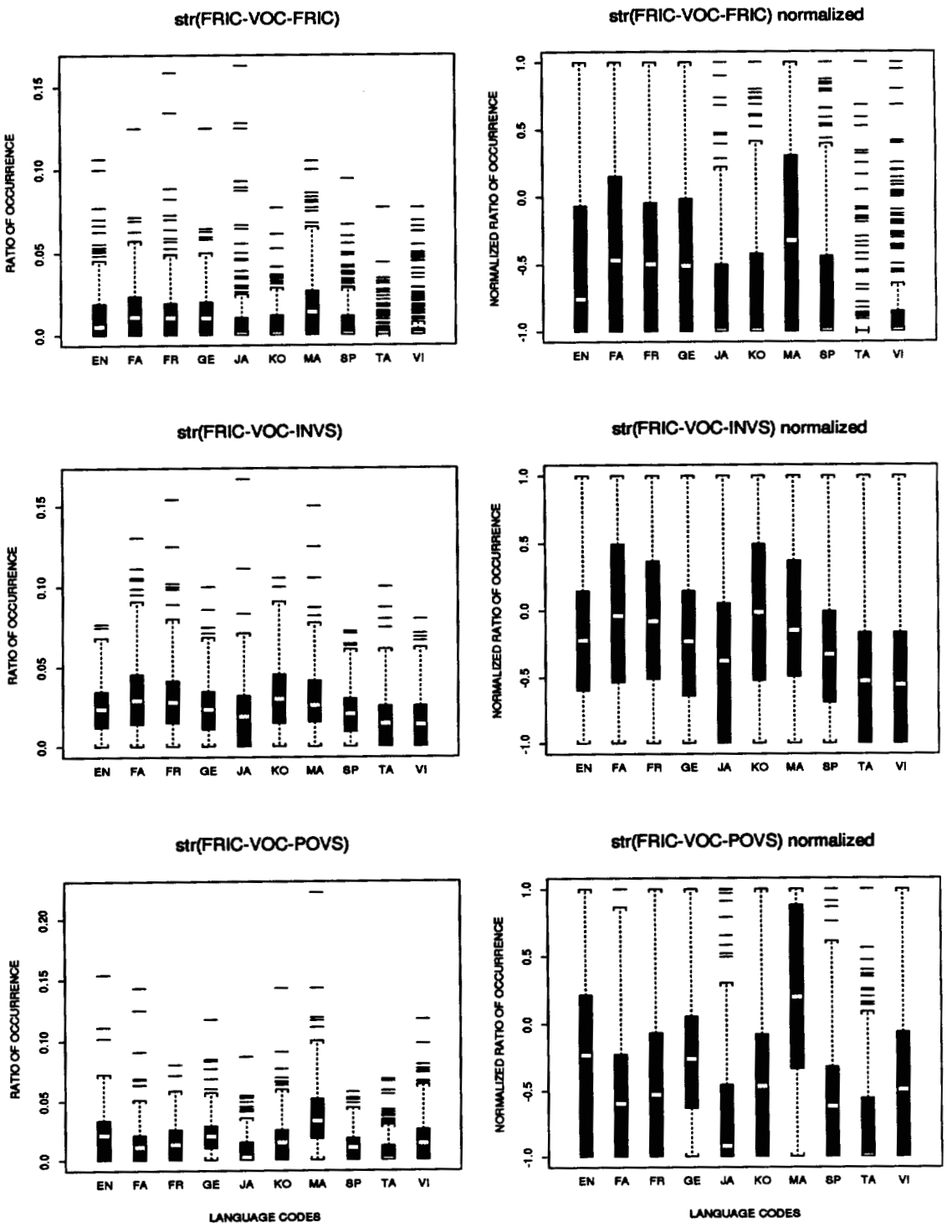

Figure C.28: Boxplots of STR Features (continued)

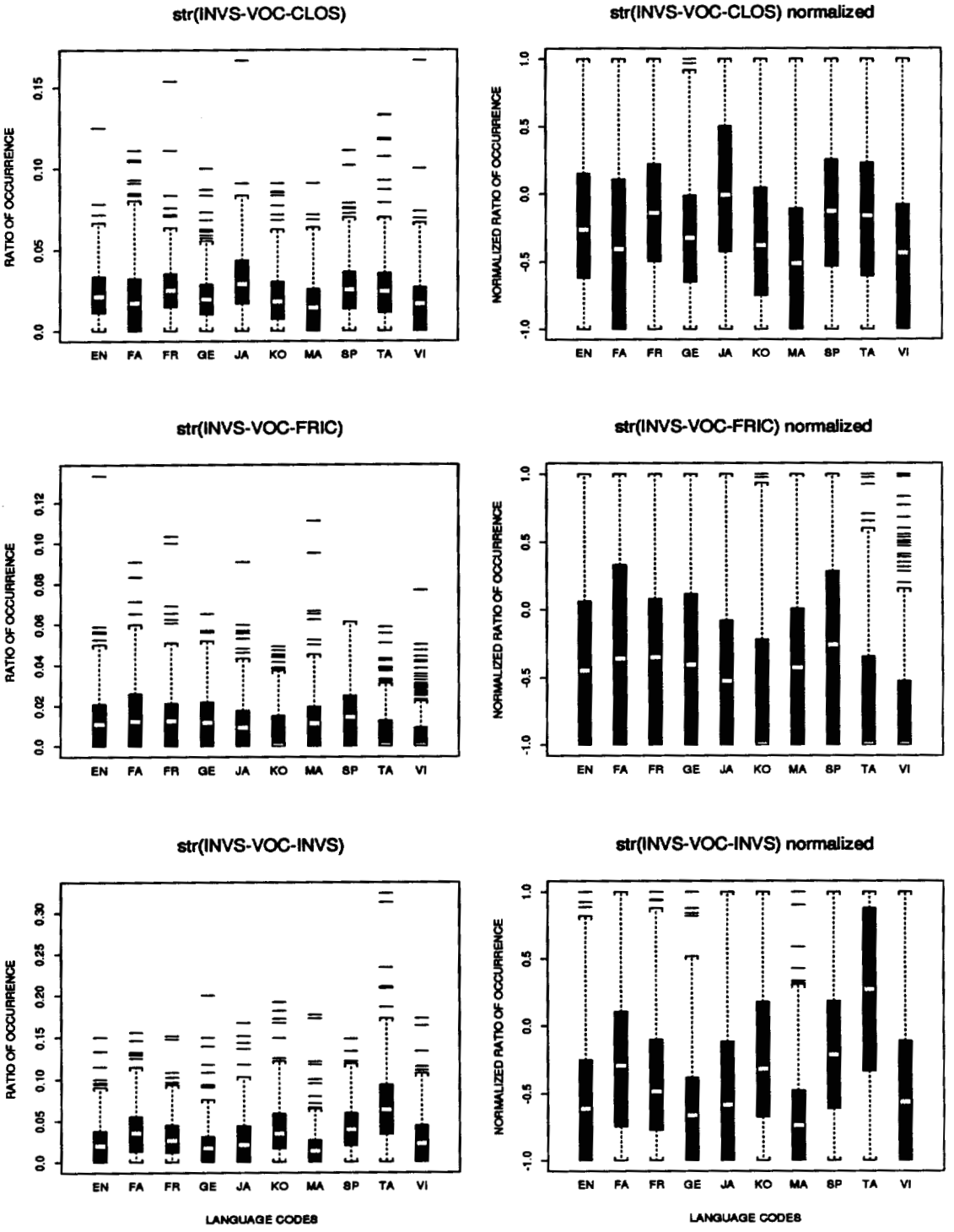

Figure C.29: Boxplots of STR Features (continued)

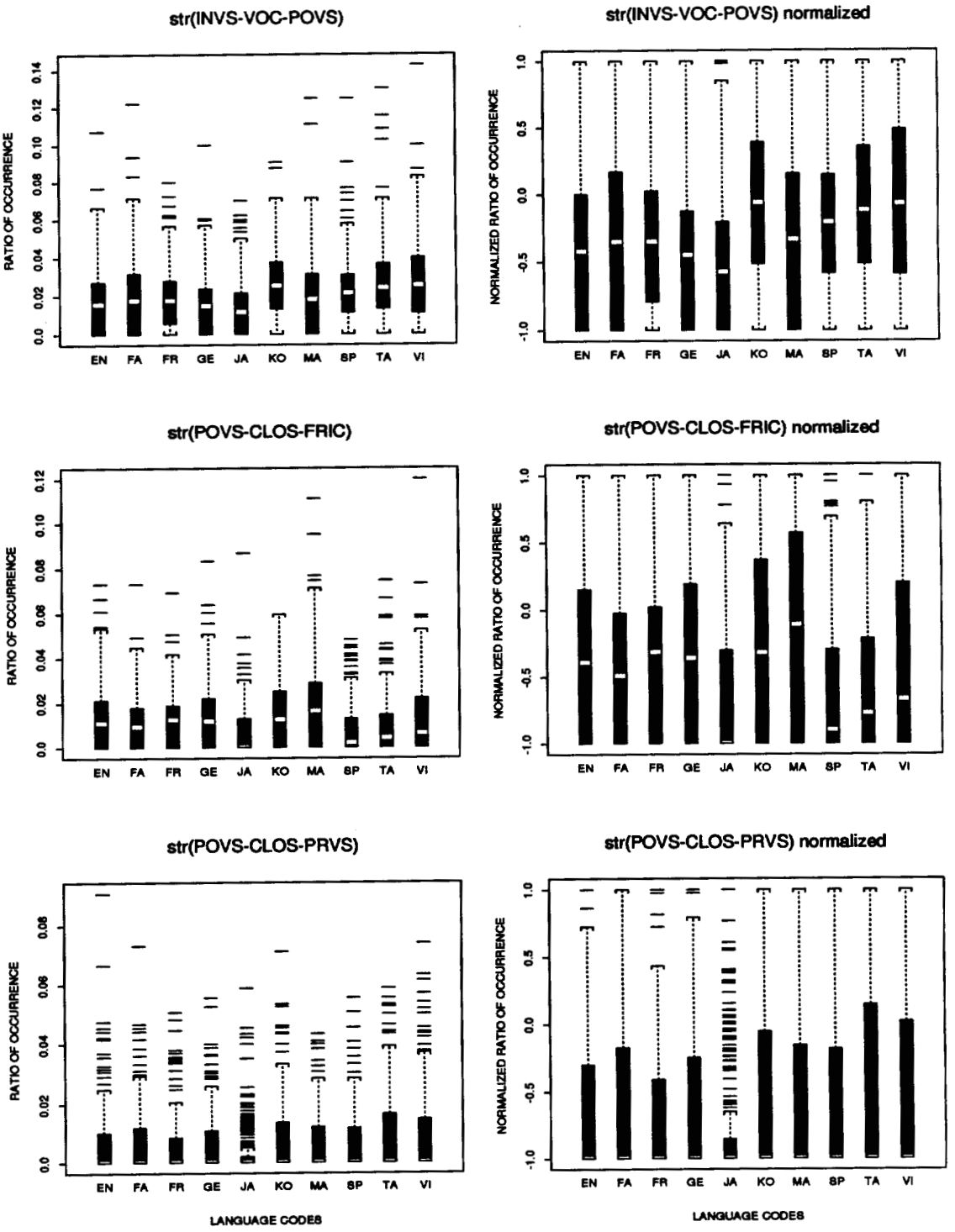

Figure C.30: Boxplots of STR Features (continued)

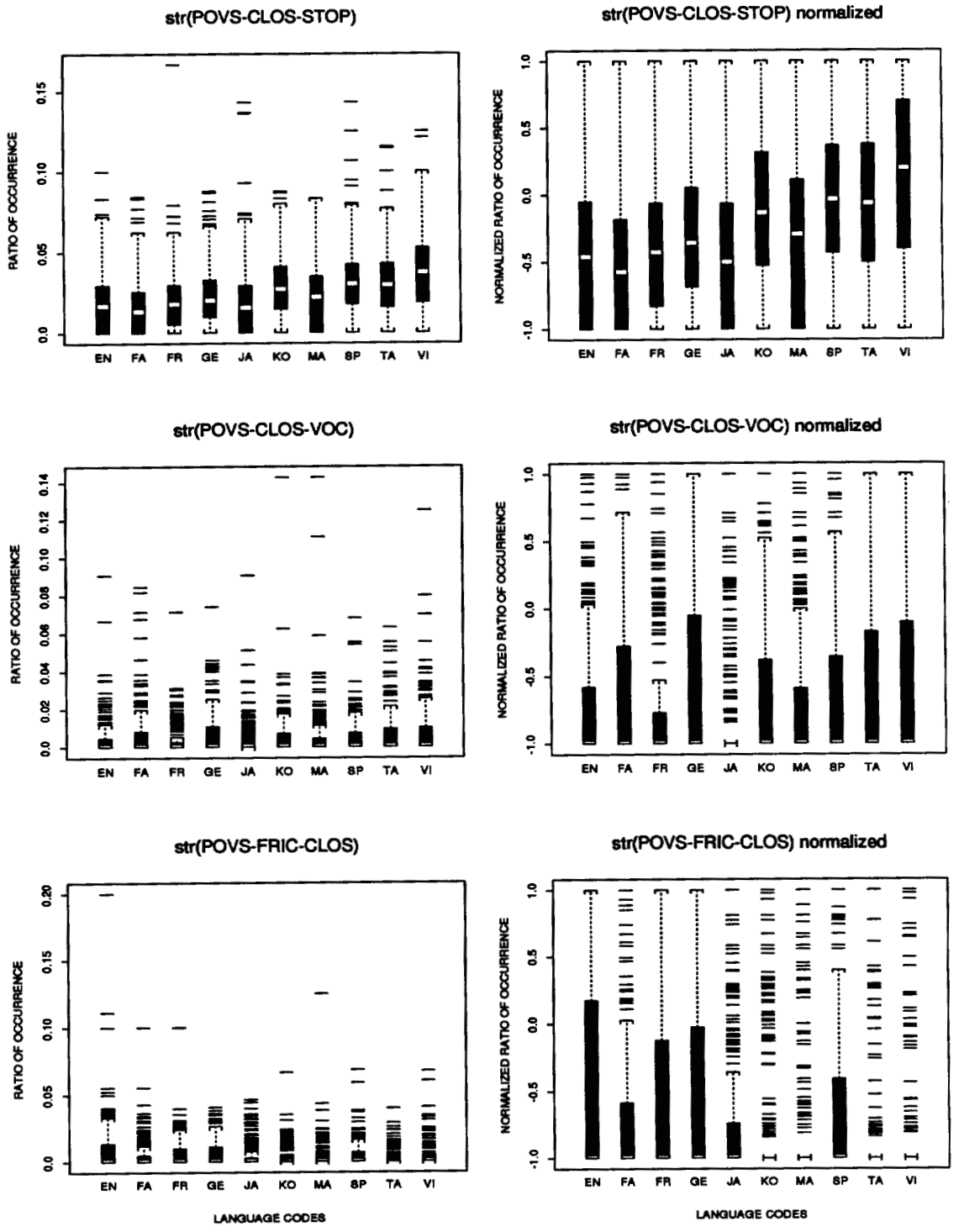

Figure C.31: Boxplots of STR Features (continued)

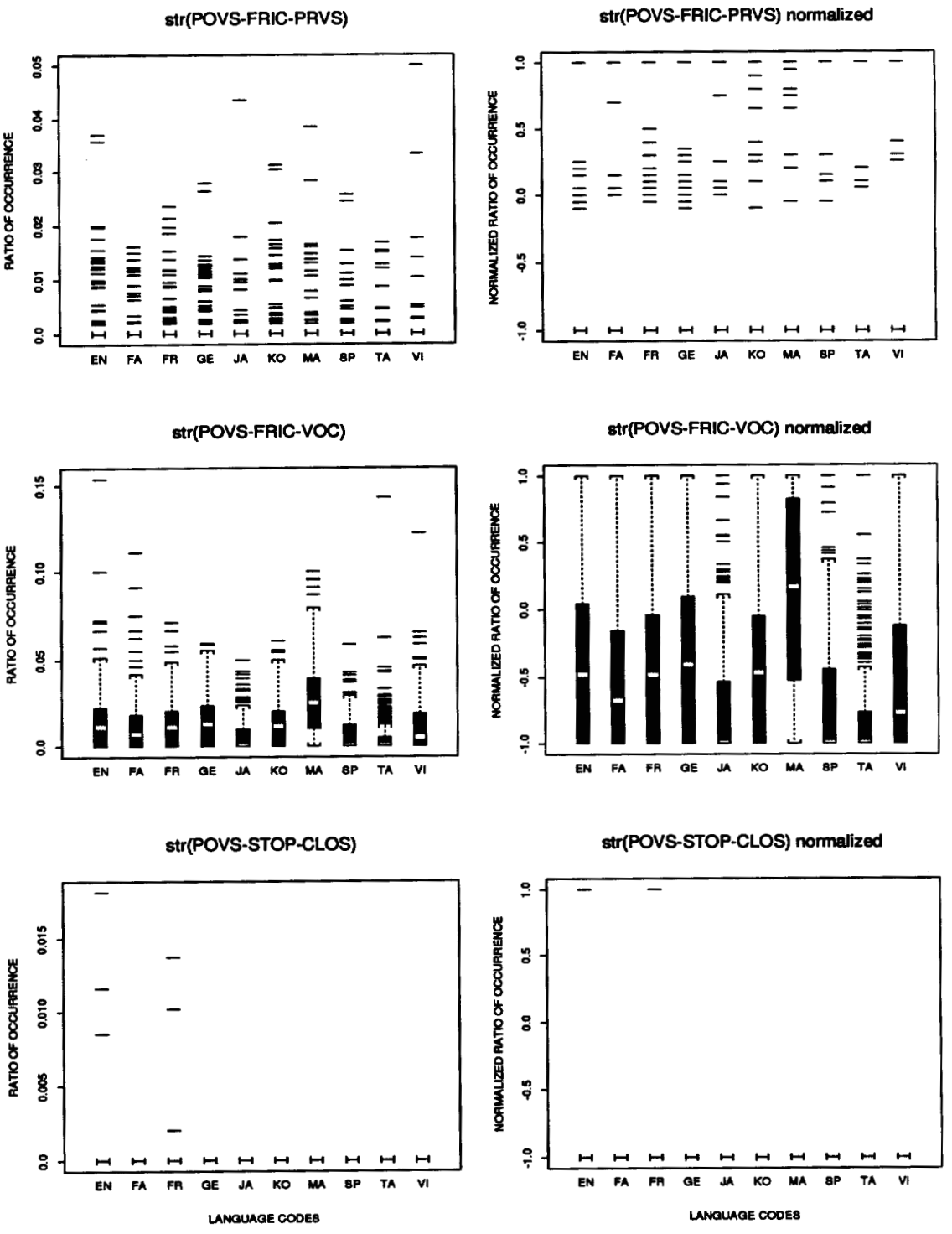

Figure C.32: Boxplots of STR Features (continued)

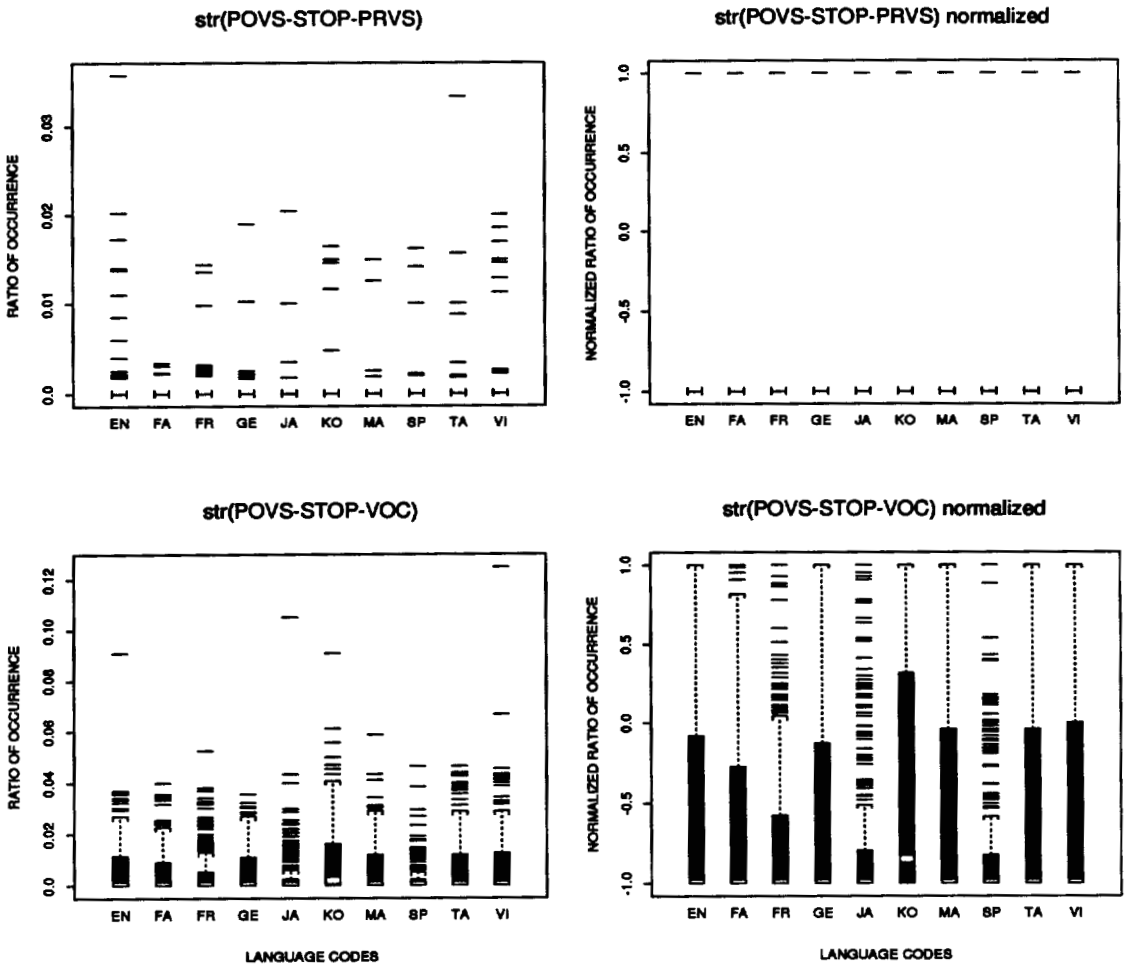

Figure C.33: Boxplots of STR Features (continued)

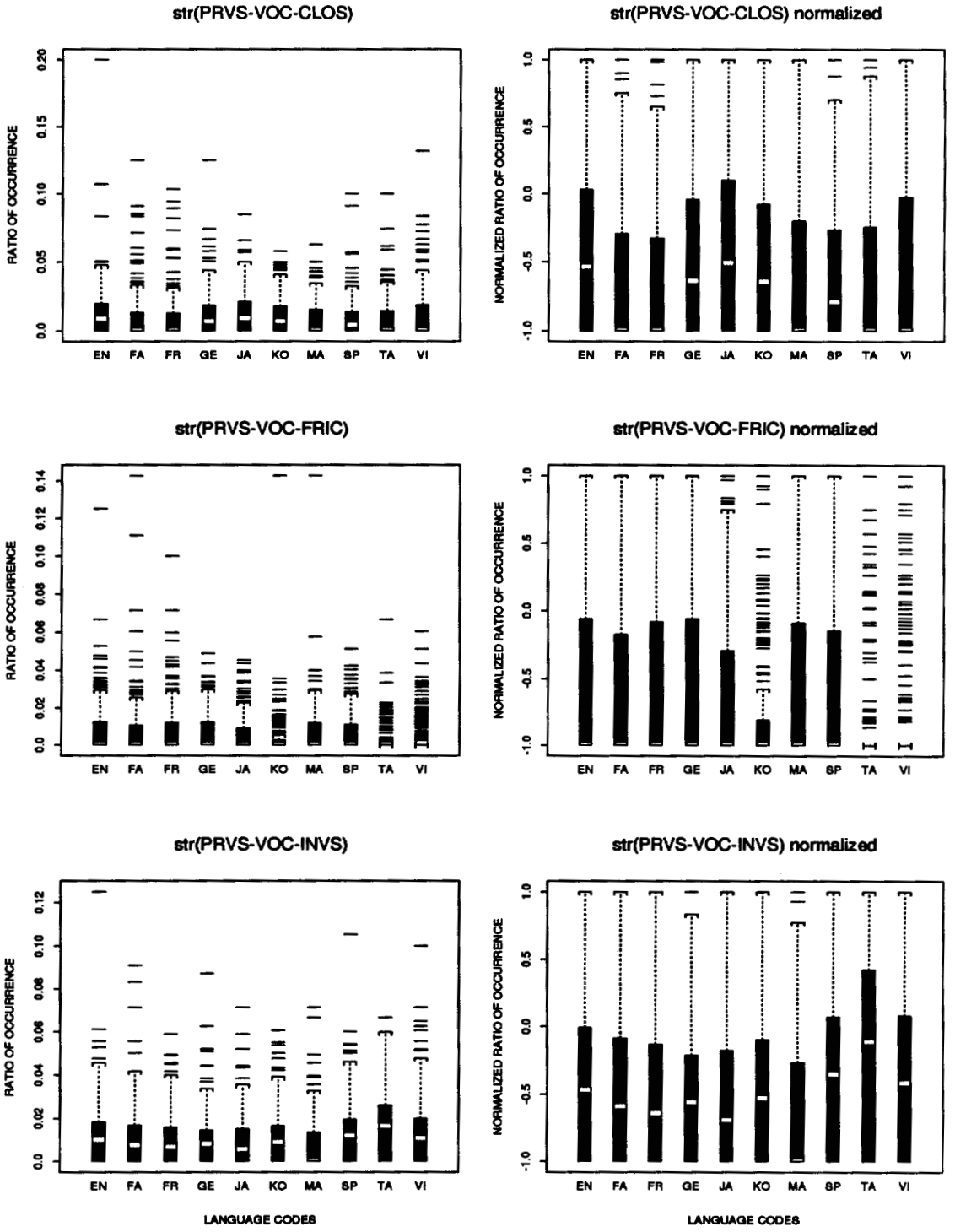

Figure C.34: Boxplots of STR Features (continued)

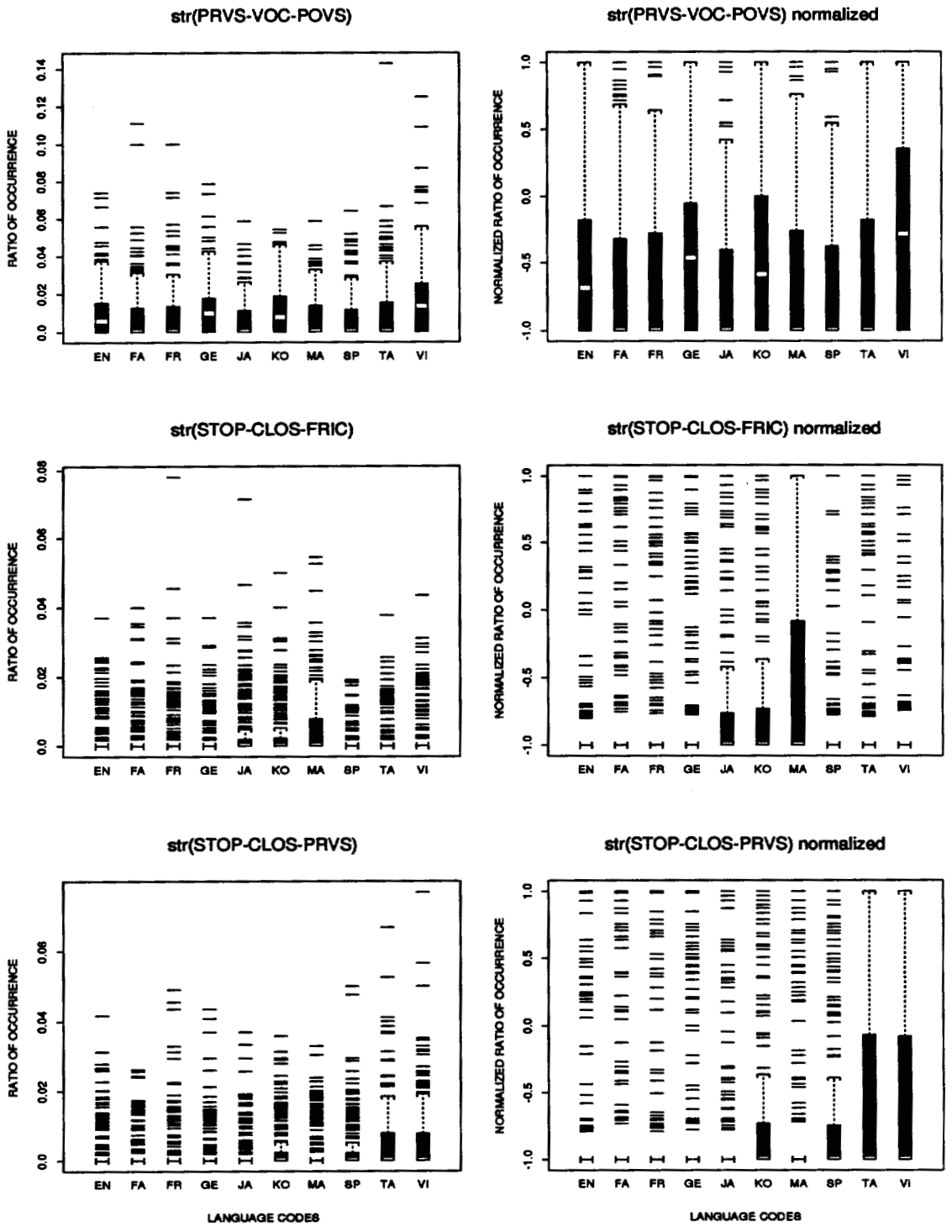

Figure C.35: Boxplots of STR Features (continued)

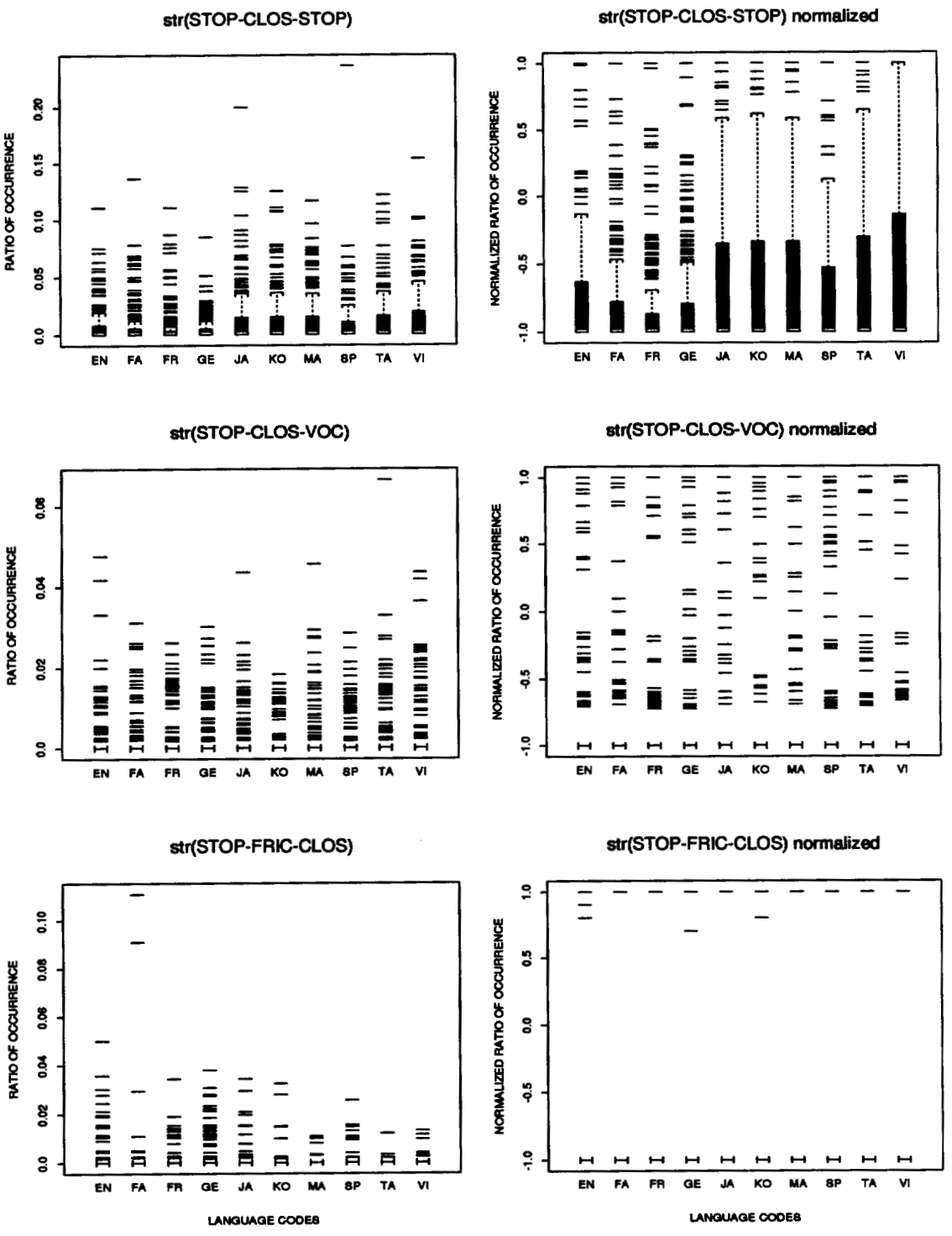

Figure C.36: Boxplots of STR Features (continued)

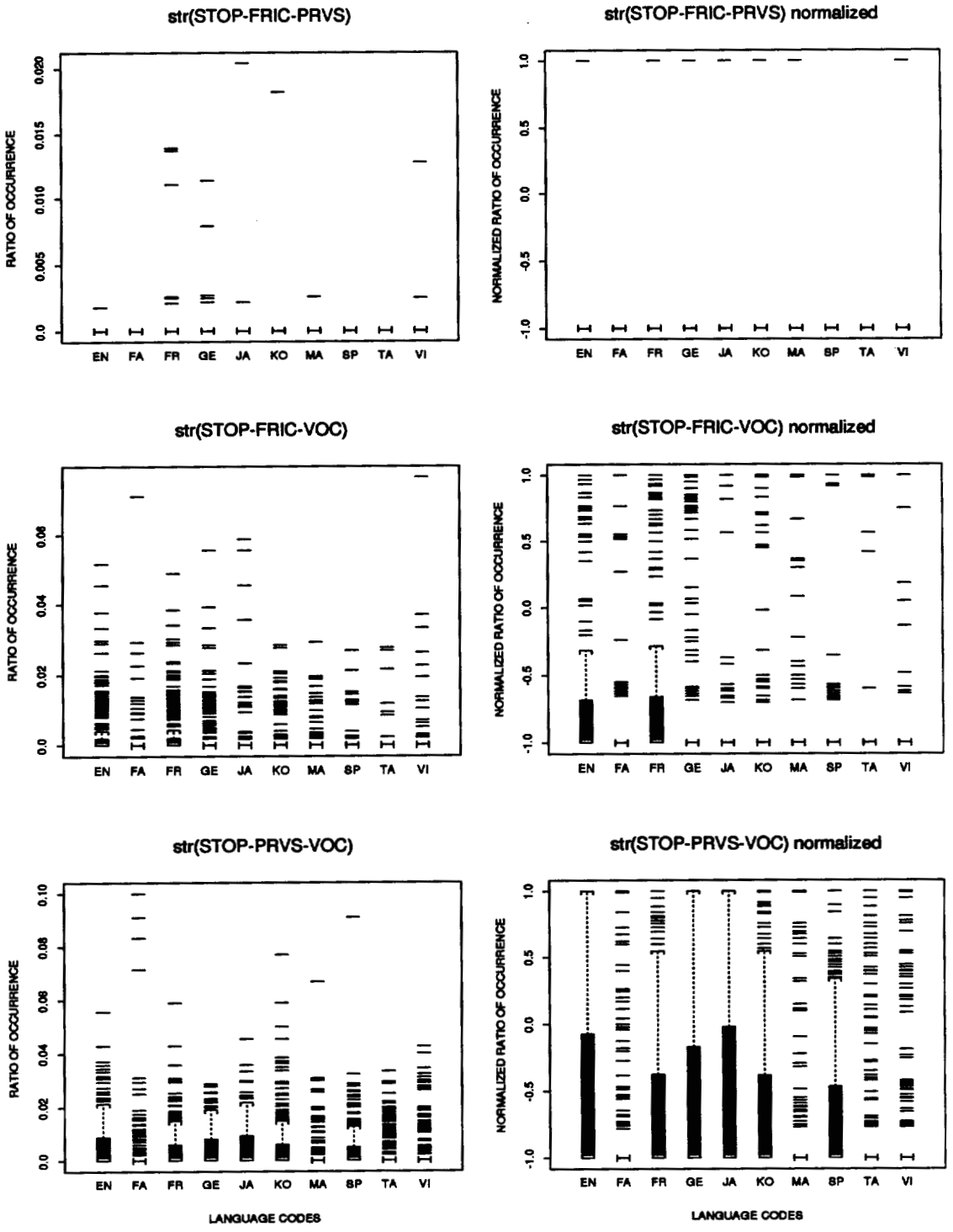

Figure C.37: Boxplots of STR Features (continued)

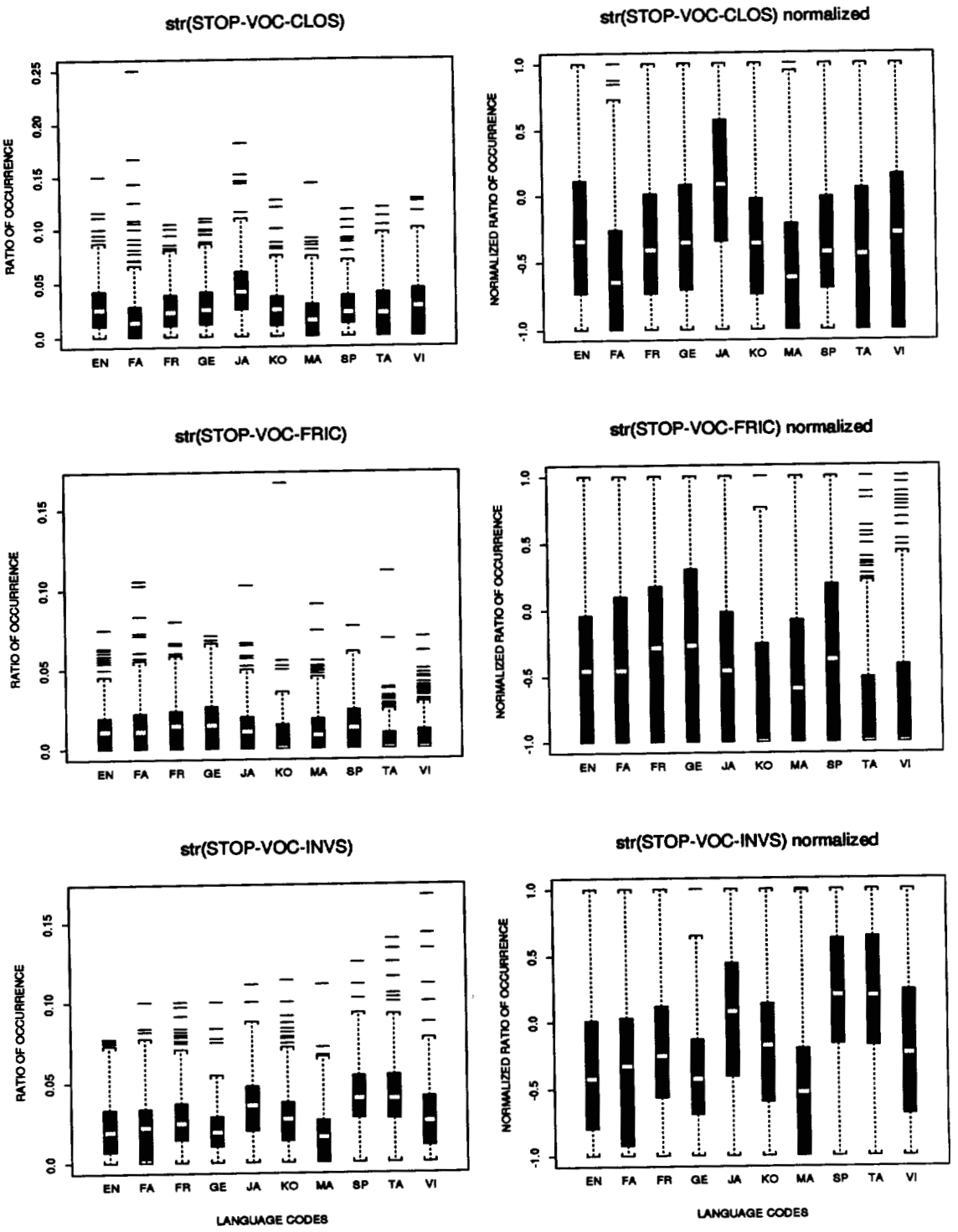

Figure C.38: Boxplots of STR Features (continued)

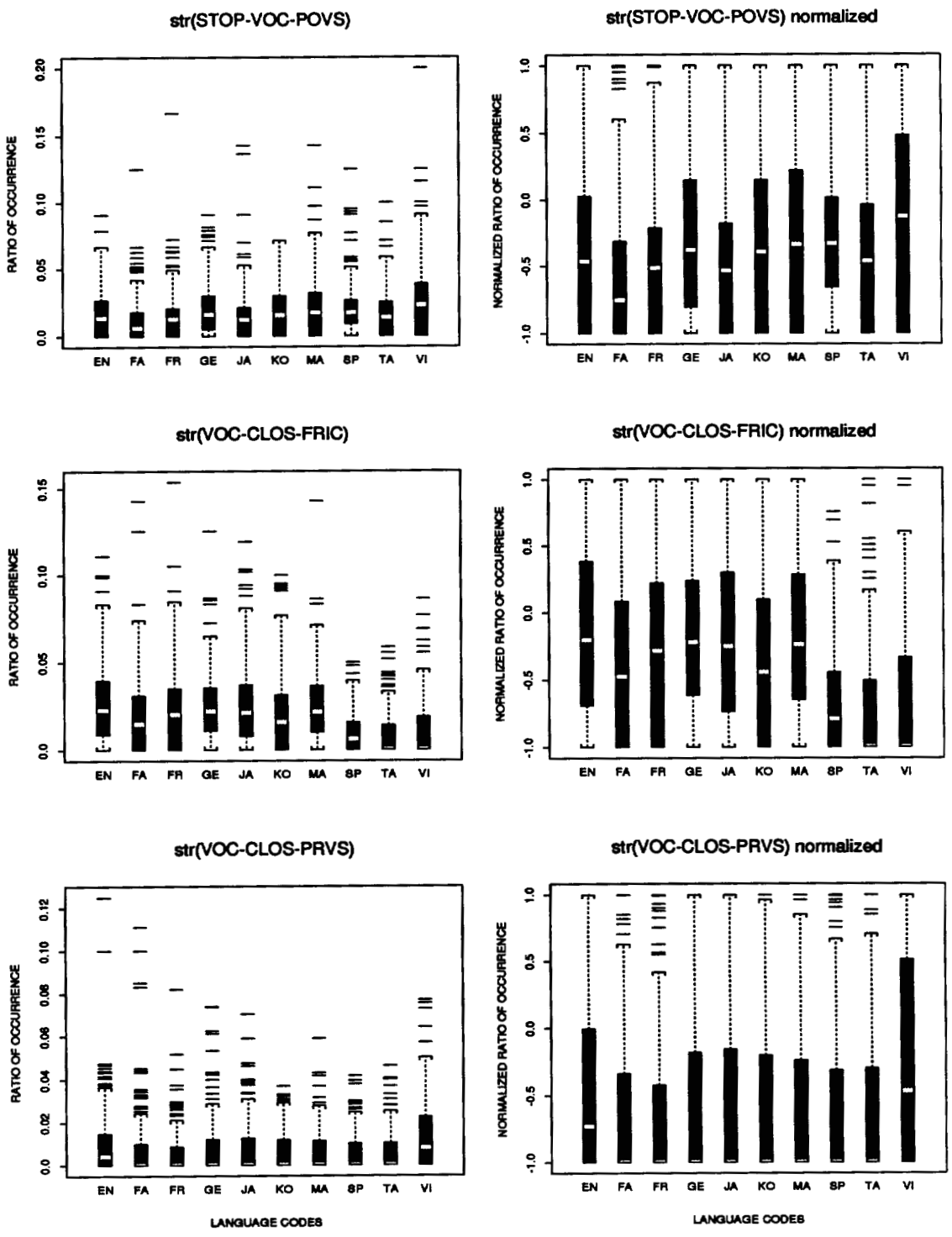

Figure C.39: Boxplots of STR Features (continued)

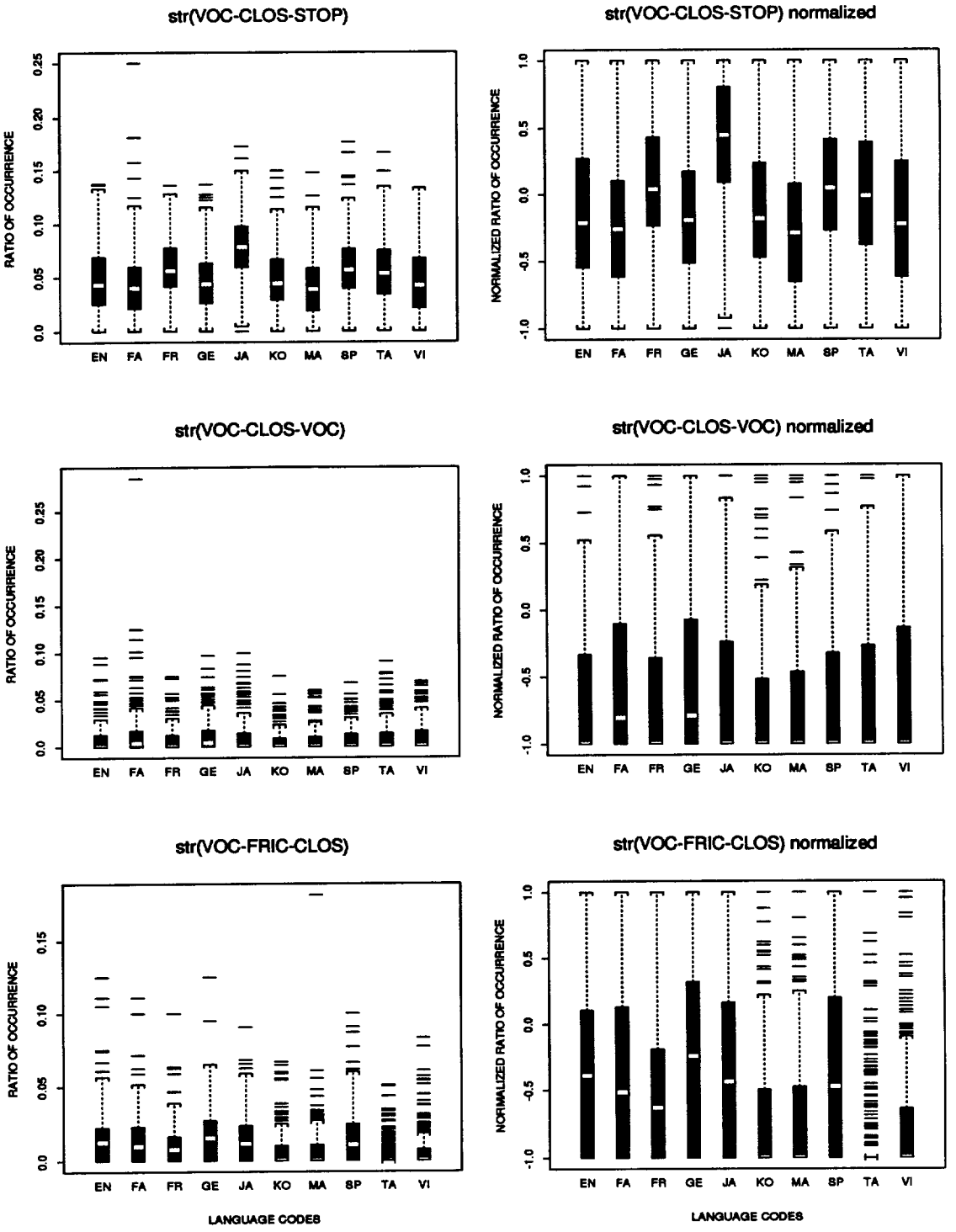

Figure C.40: Boxplots of STR Features (continued)

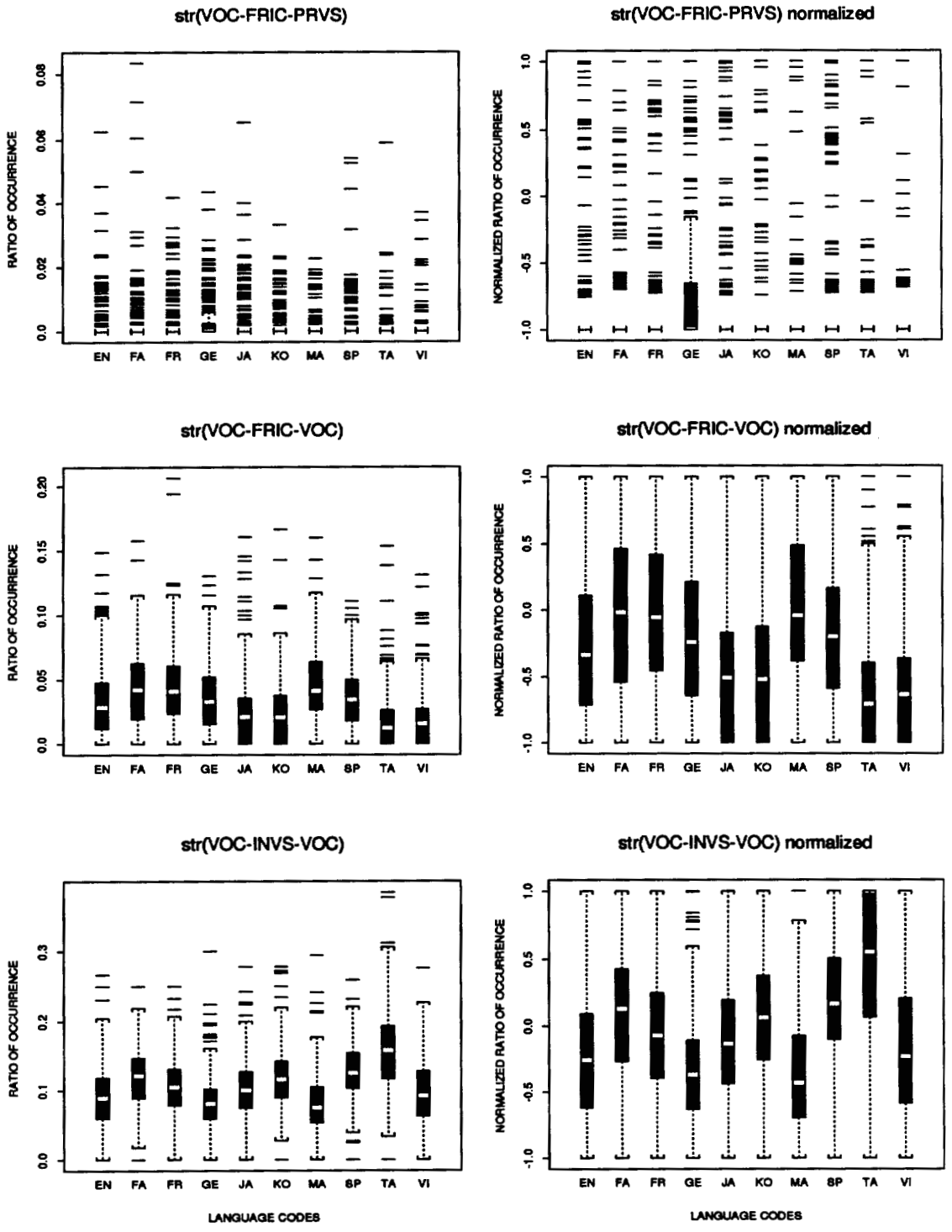

Figure C.41: Boxplots of STR Features (continued)

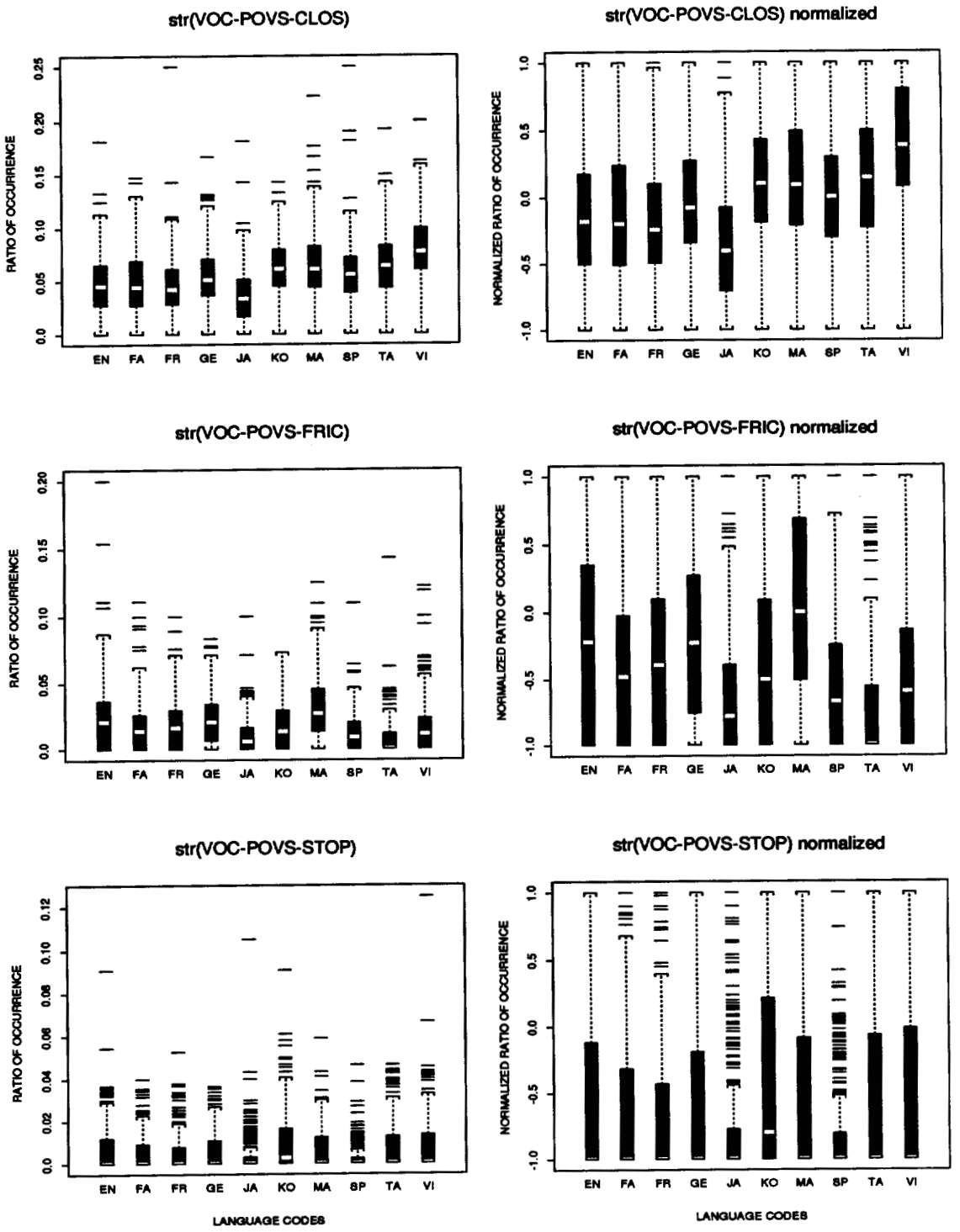

Figure C.42: Boxplots of STR Features (continued)

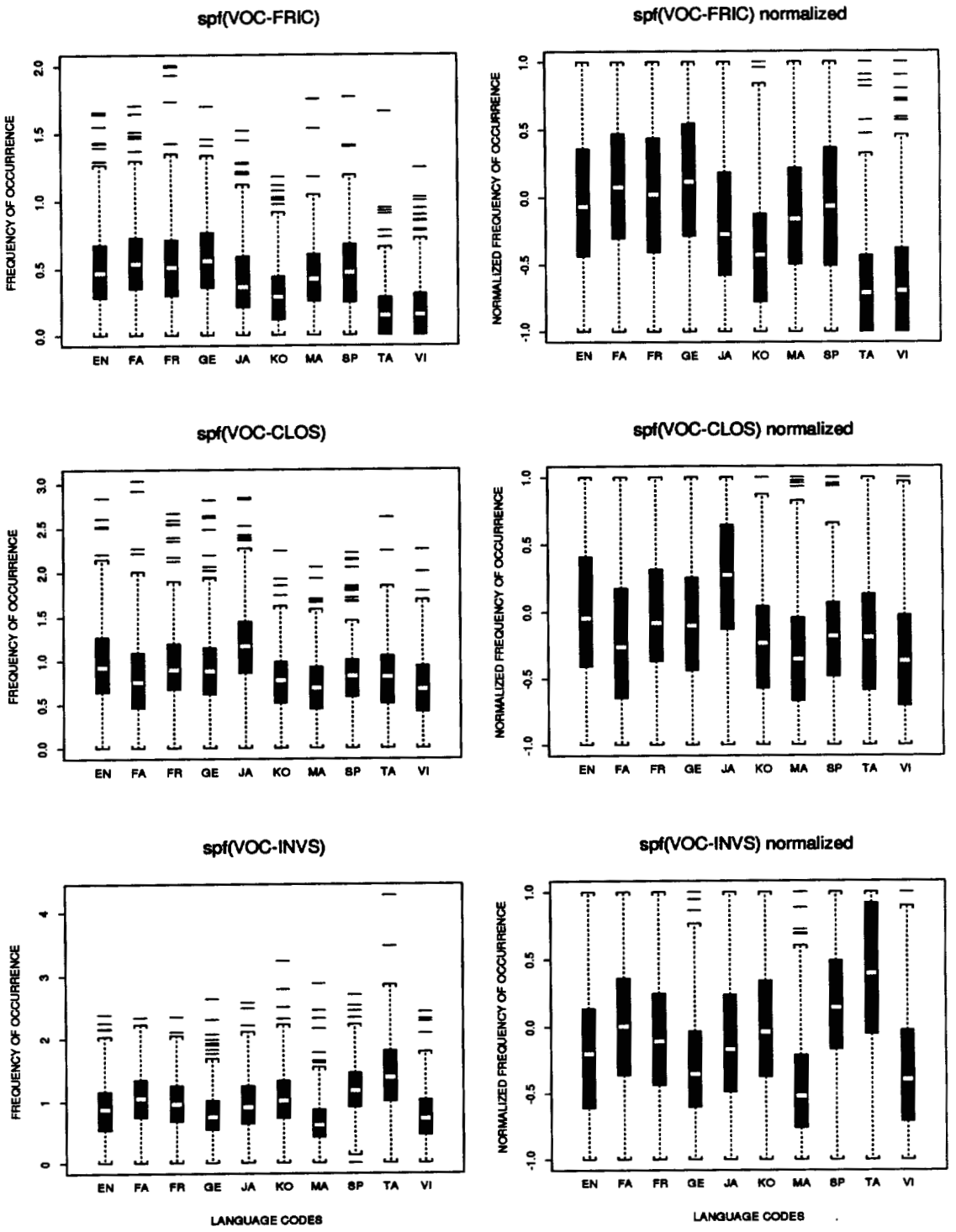

Figure C.43: Boxplots of Segment-pair Frequency (SPF) Features

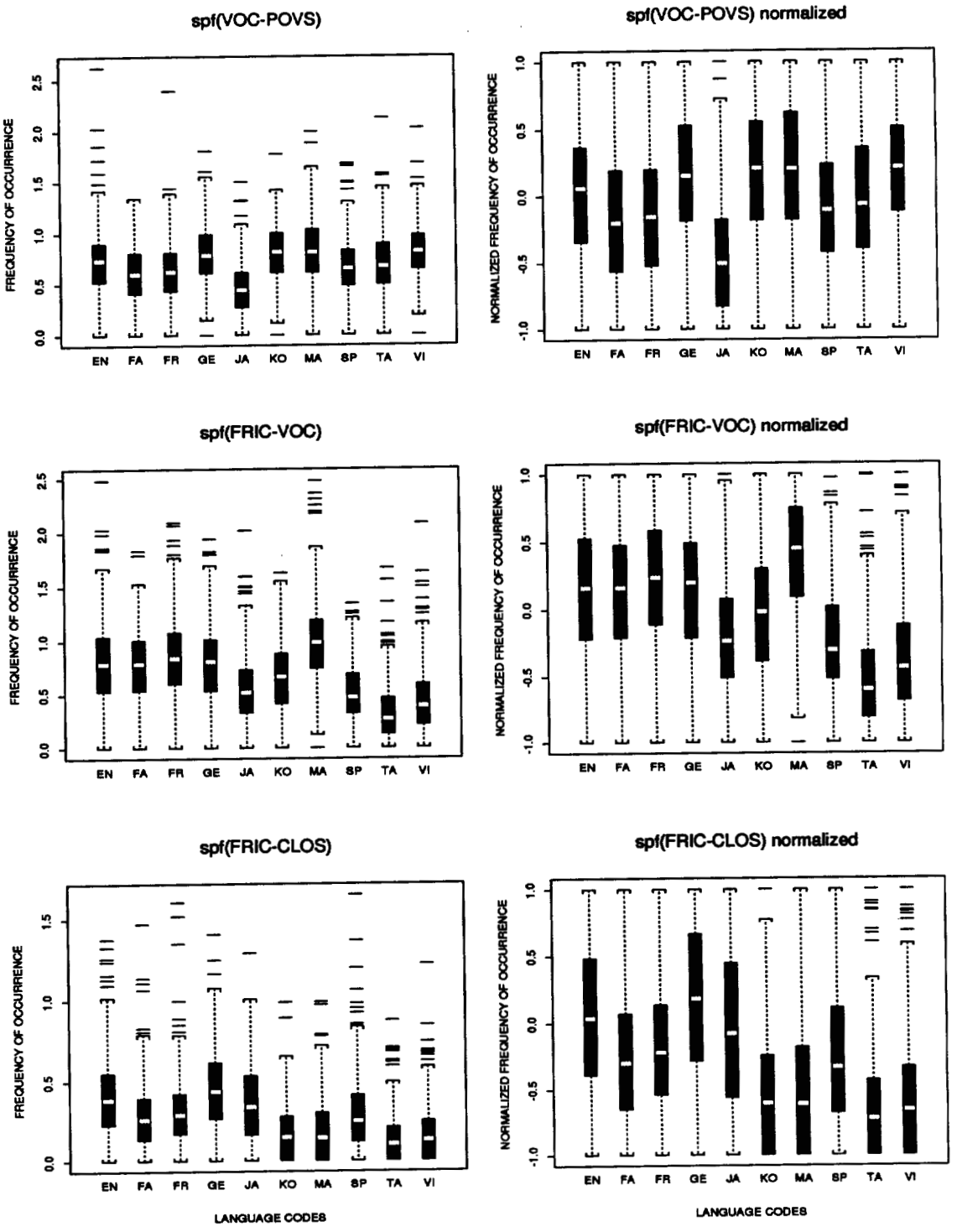

Figure C.44: Boxplots of SPF Features (continued)

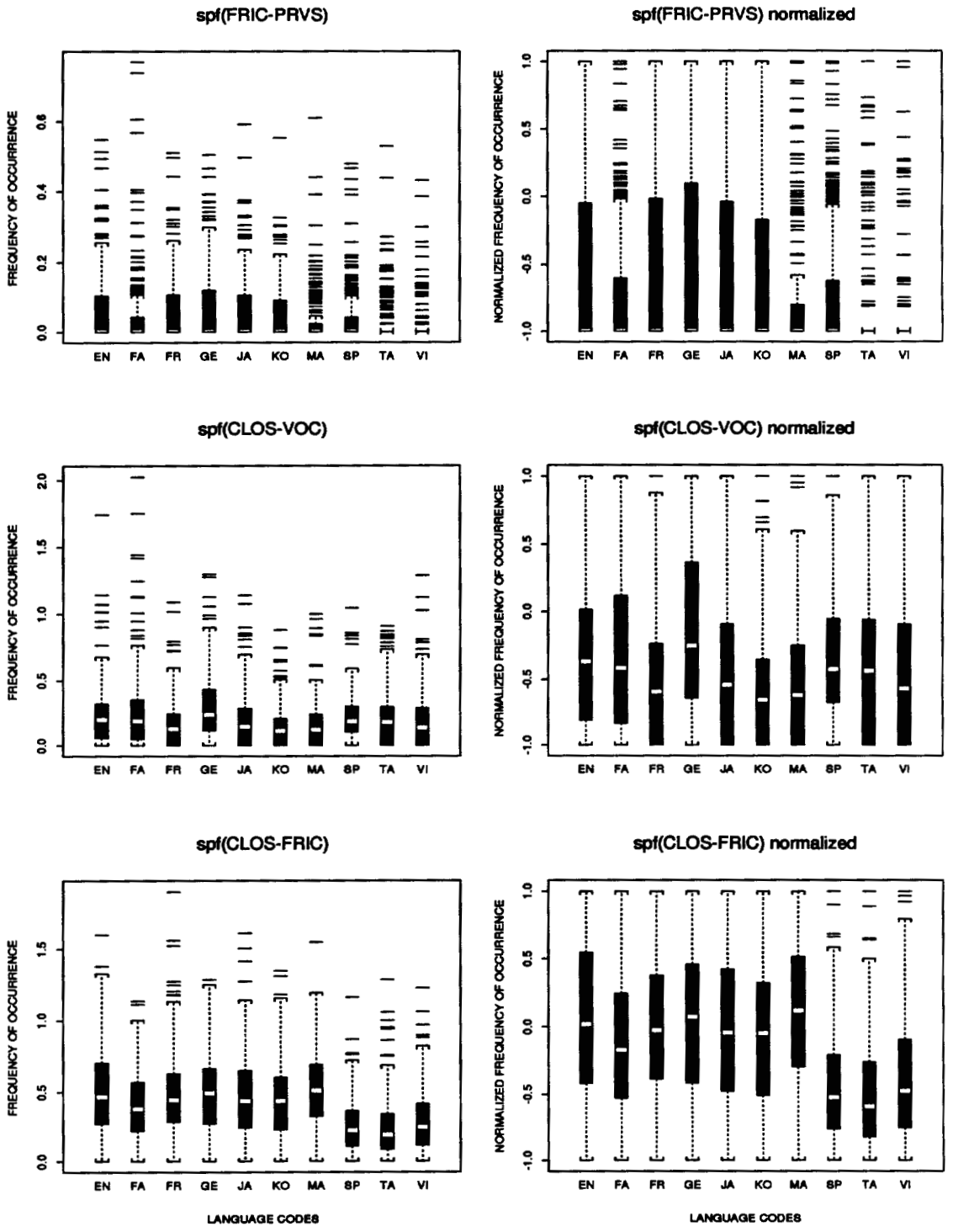

Figure C.45: Boxplots of SPF Features (continued)

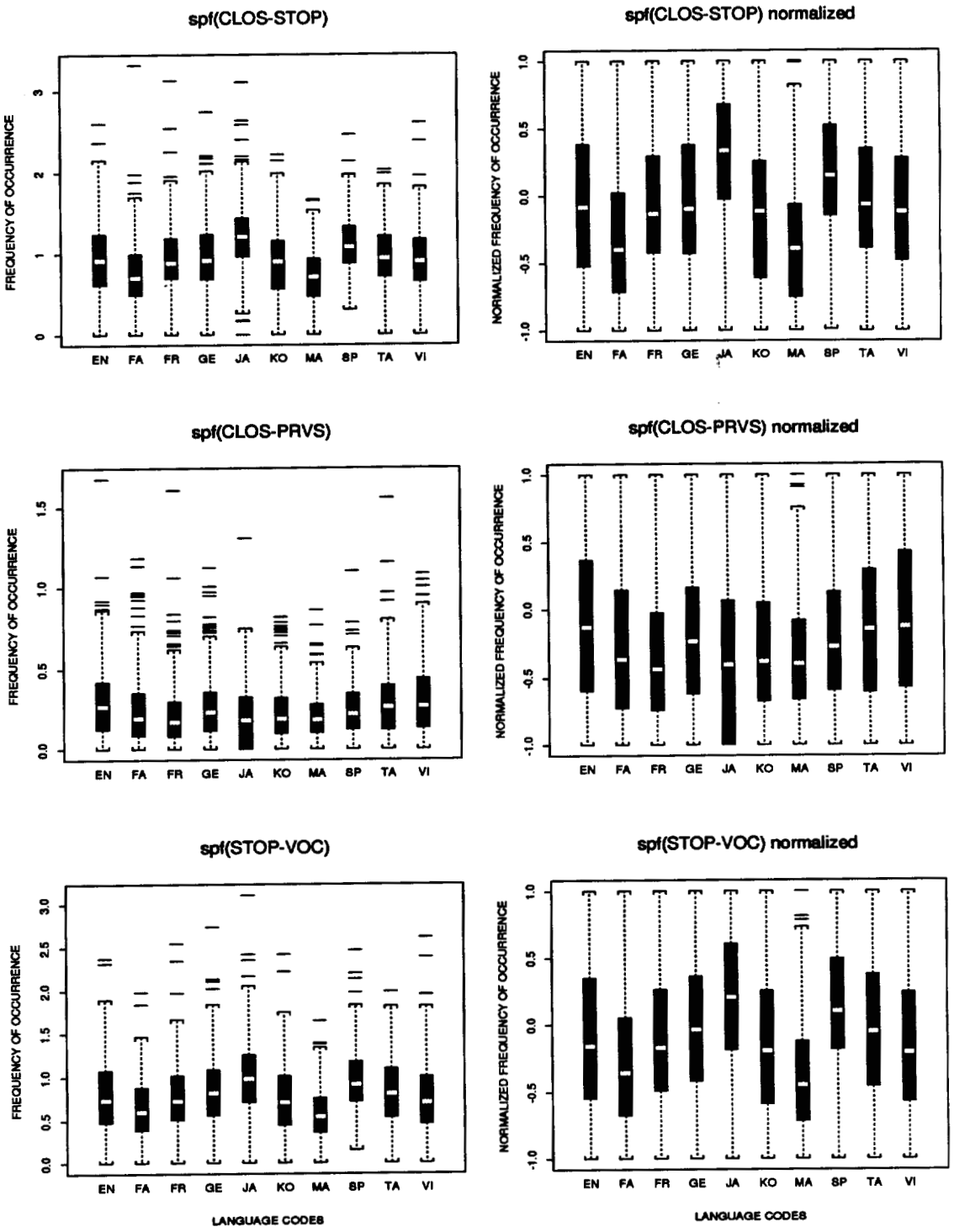

Figure C.46: Boxplots of SPF Features (continued)

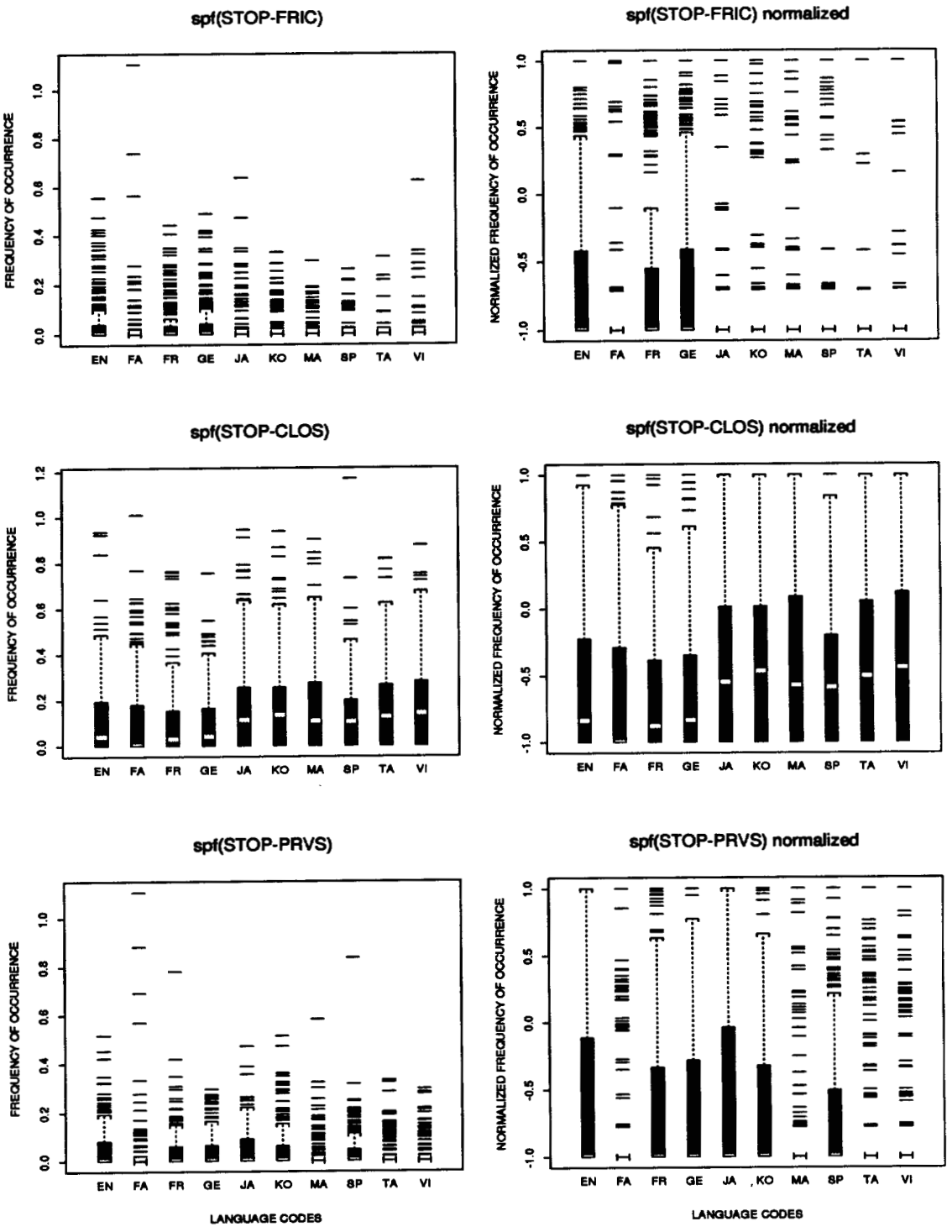

Figure C.47: Boxplots of SPF Features (continued)

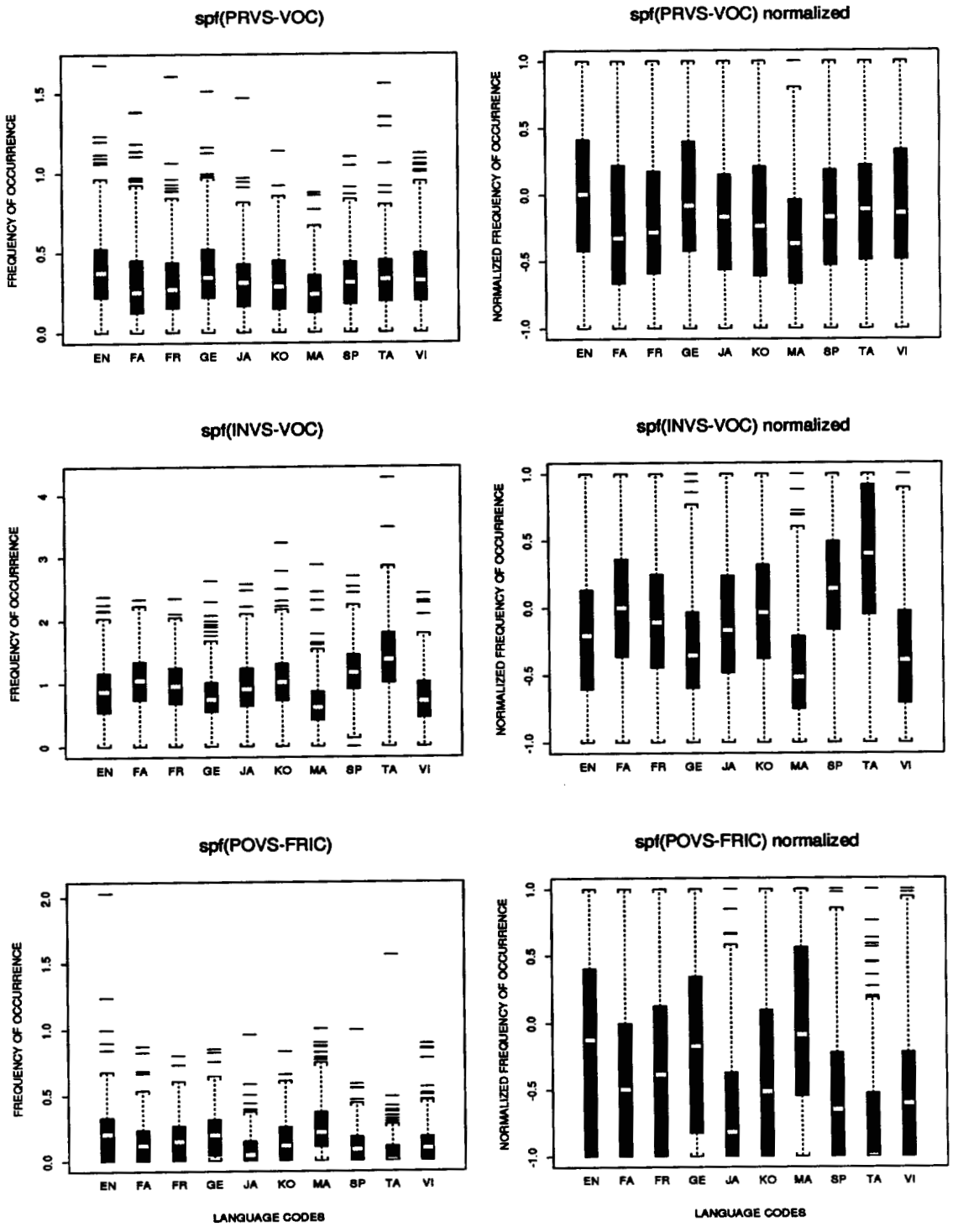

Figure C.48: Boxplots of SPF Features (continued)

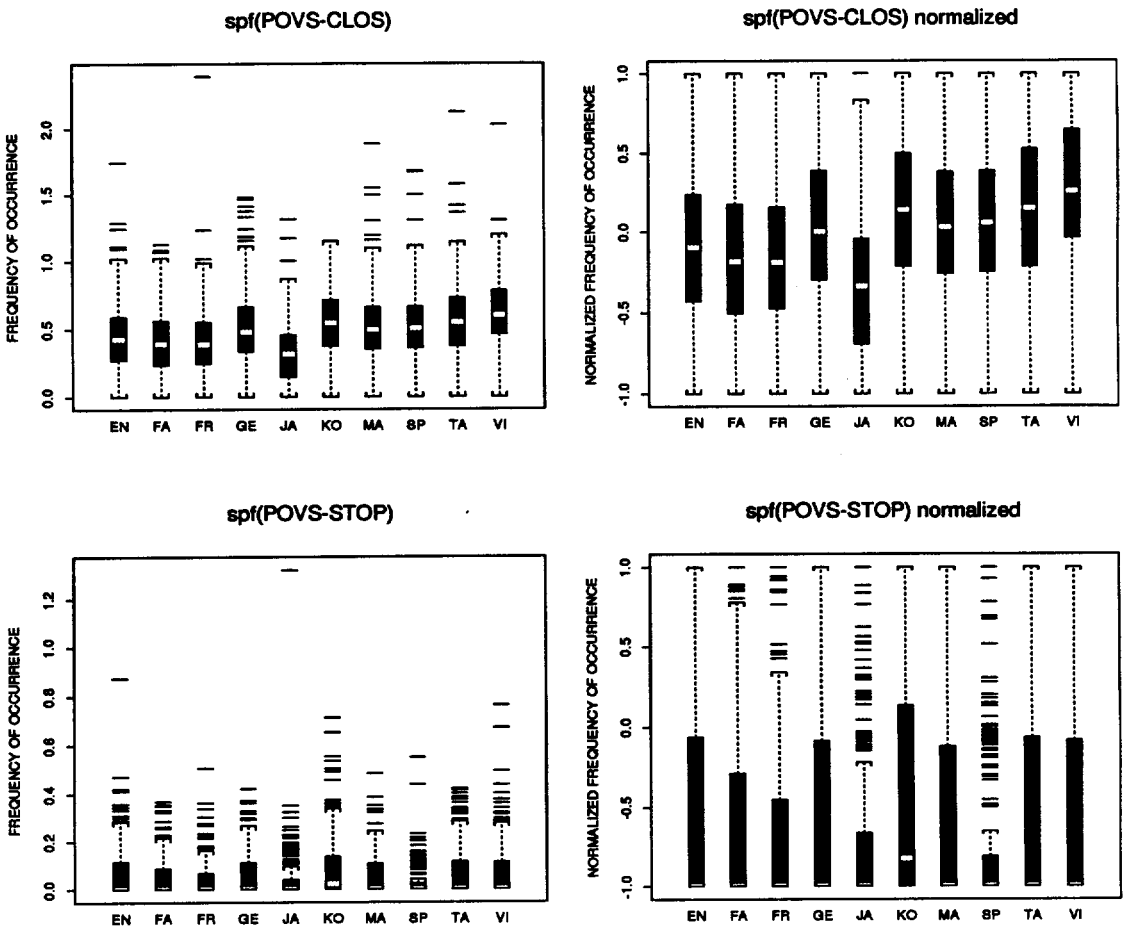

Figure C.49: Boxplots of SPF Features (continued)

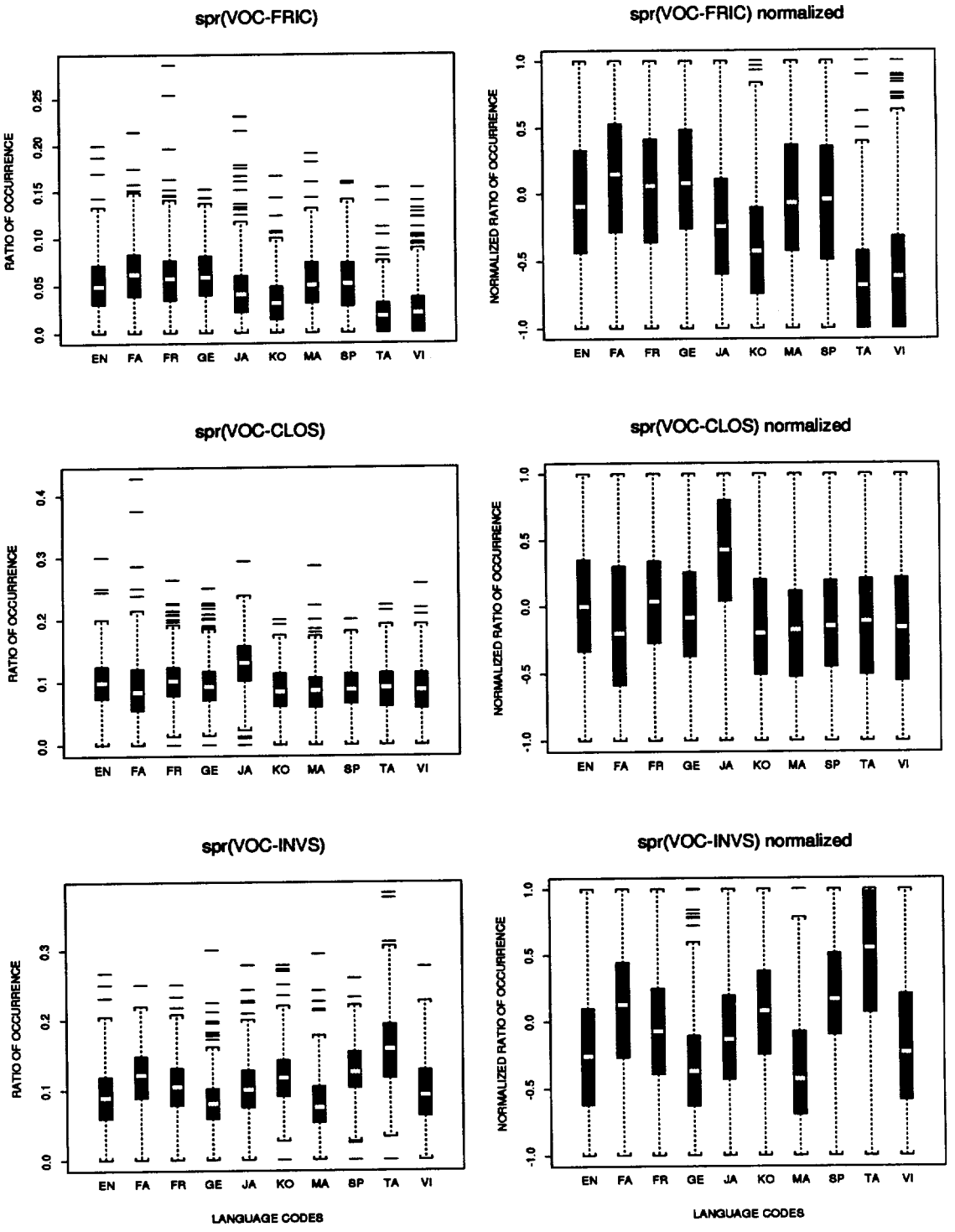

Figure C.50: Boxplots of Segment-pair Ratio (SPR) Features

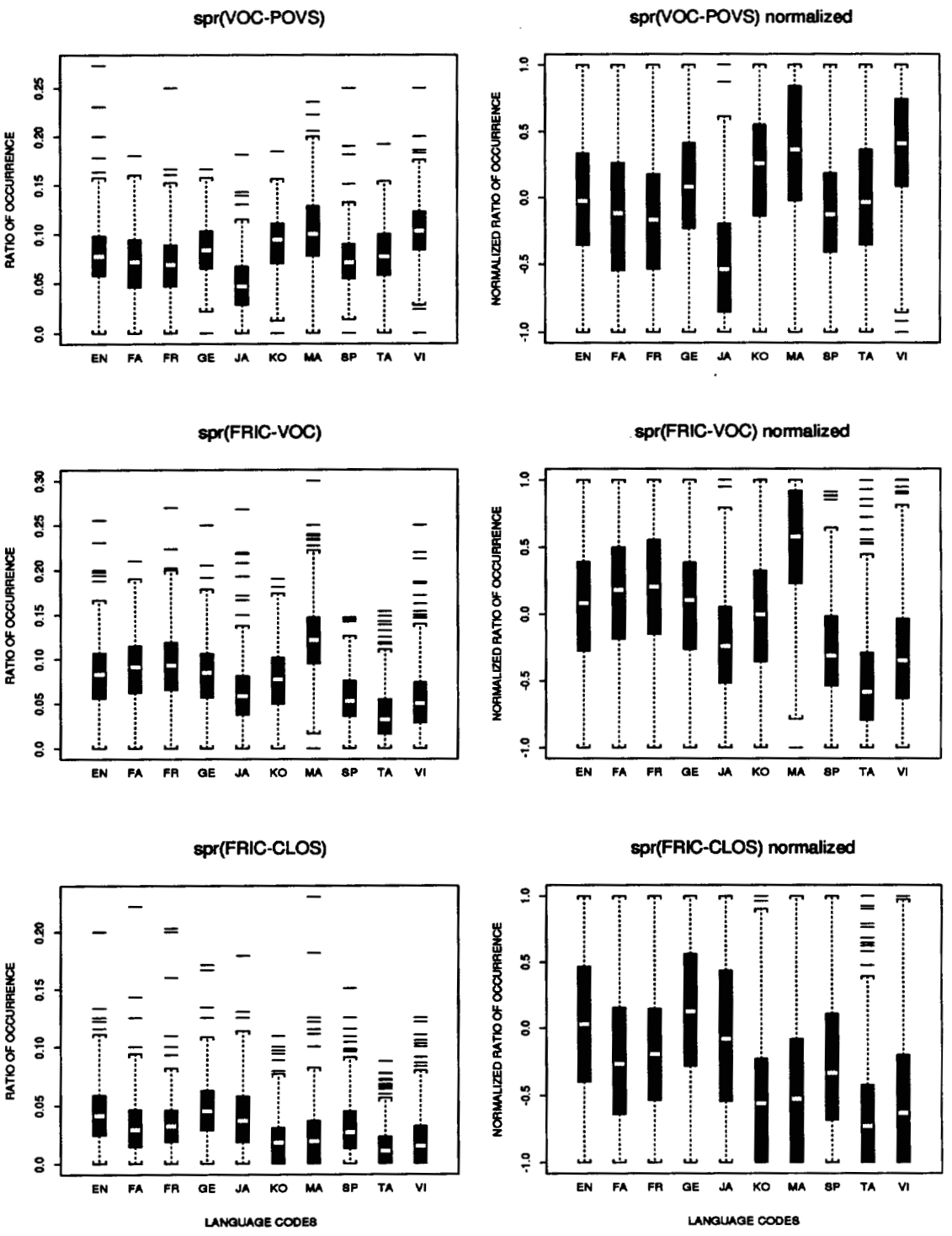

Figure C.51: Boxplots of SPR Features (continued)

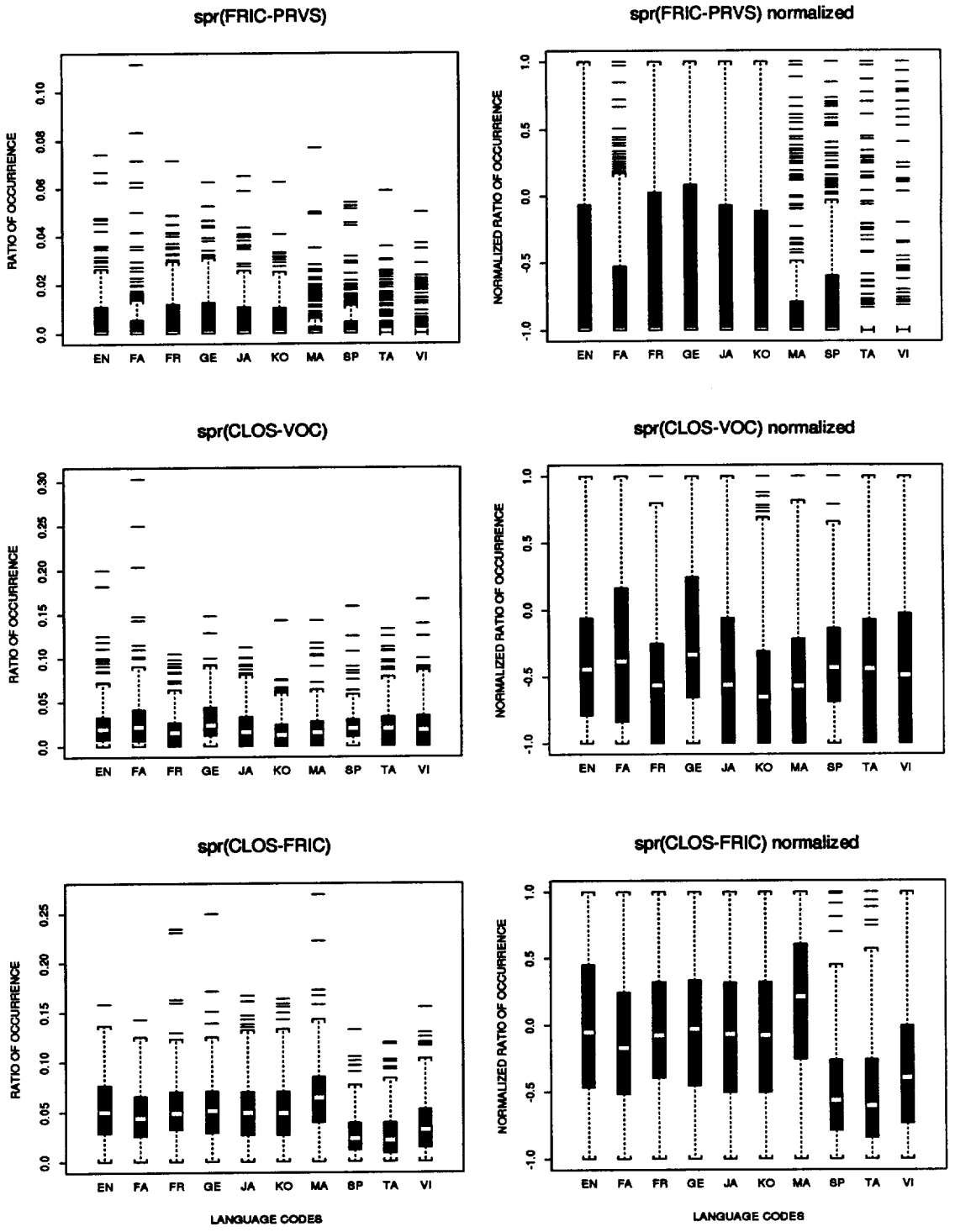

Figure C.52: Boxplots of SPR Features (continued)

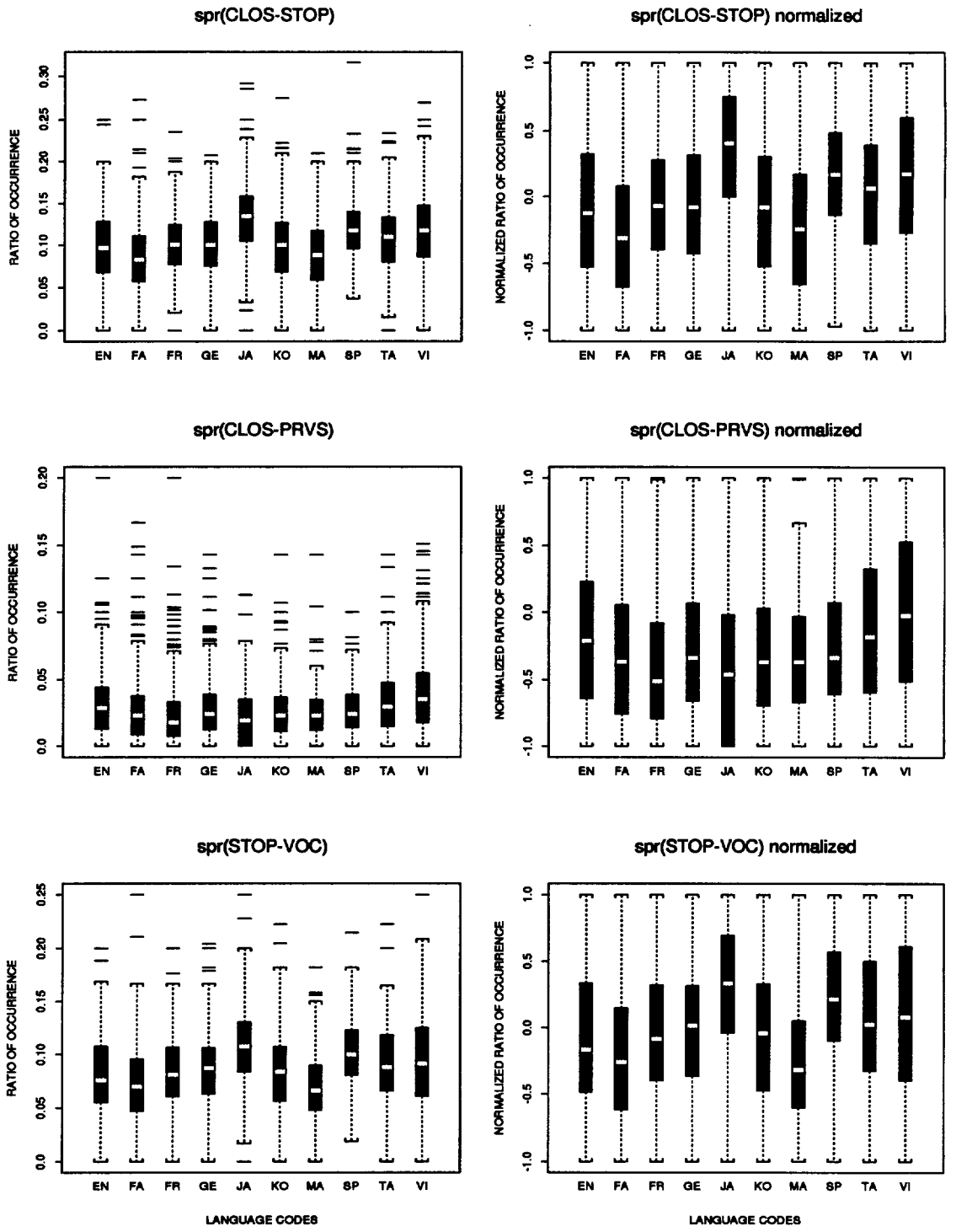

Figure C.53: Boxplots of SPR Features (continued)

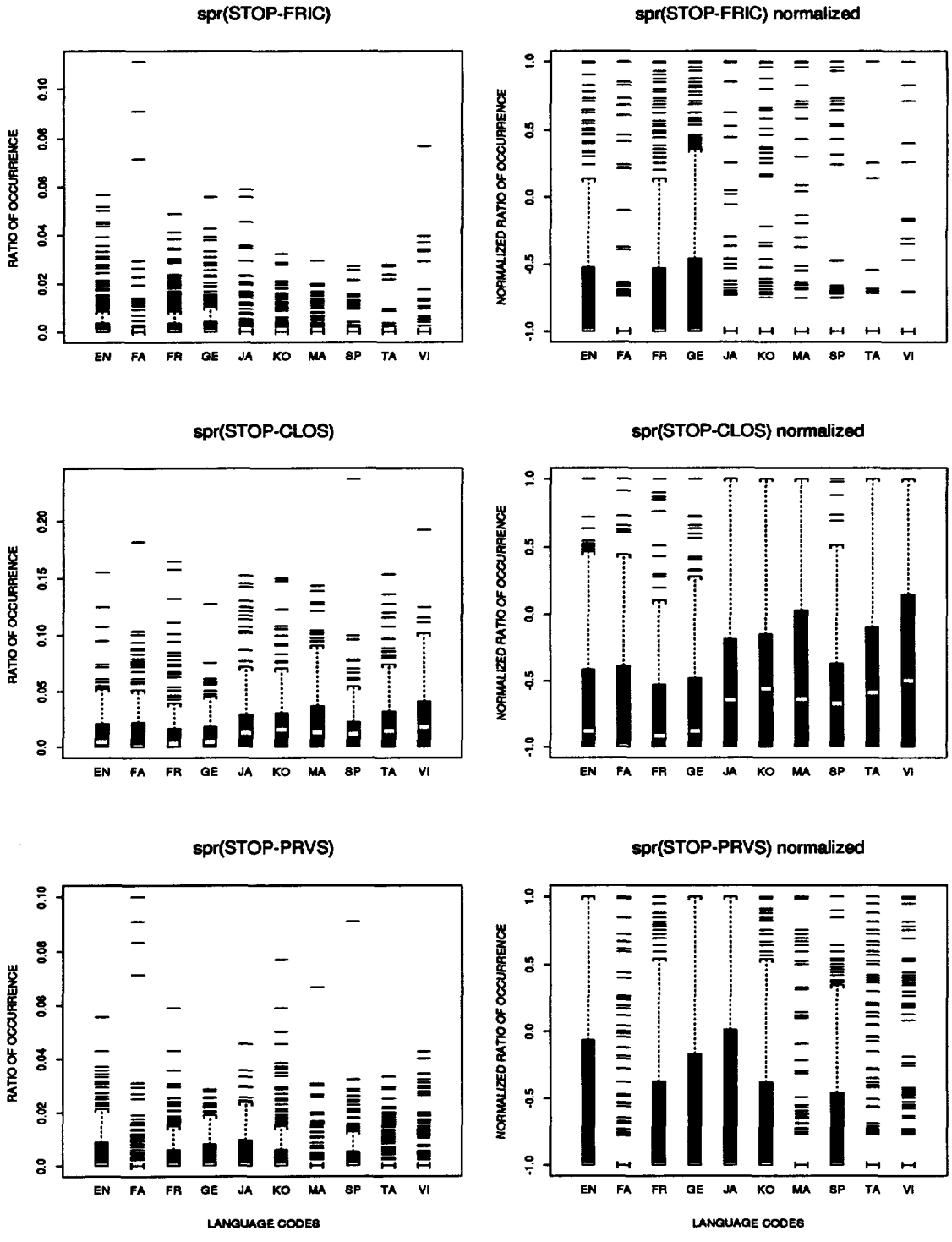

Figure C.54: Boxplots of SPR Features (continued)

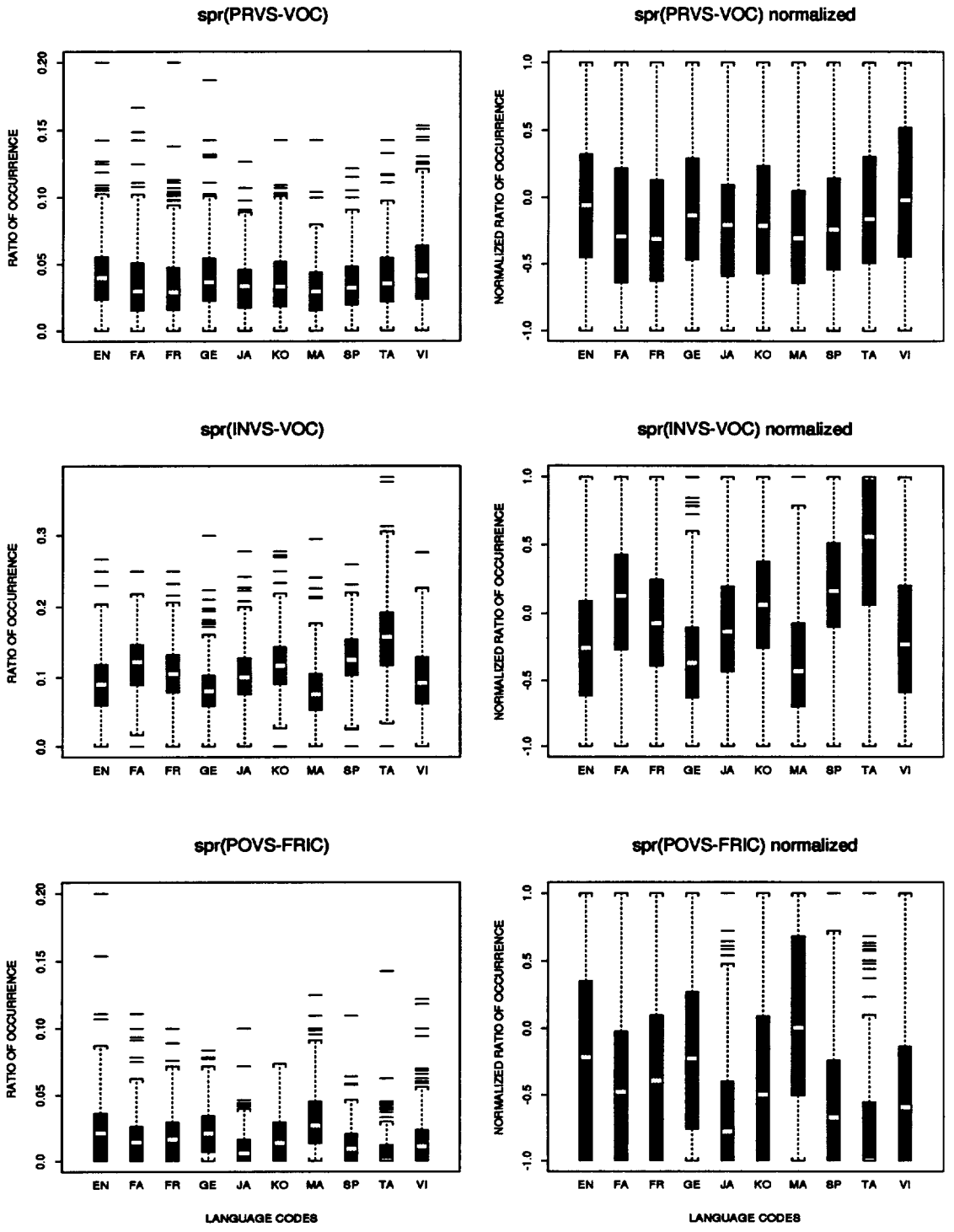

Figure C.55: Boxplots of SPR Features (continued)

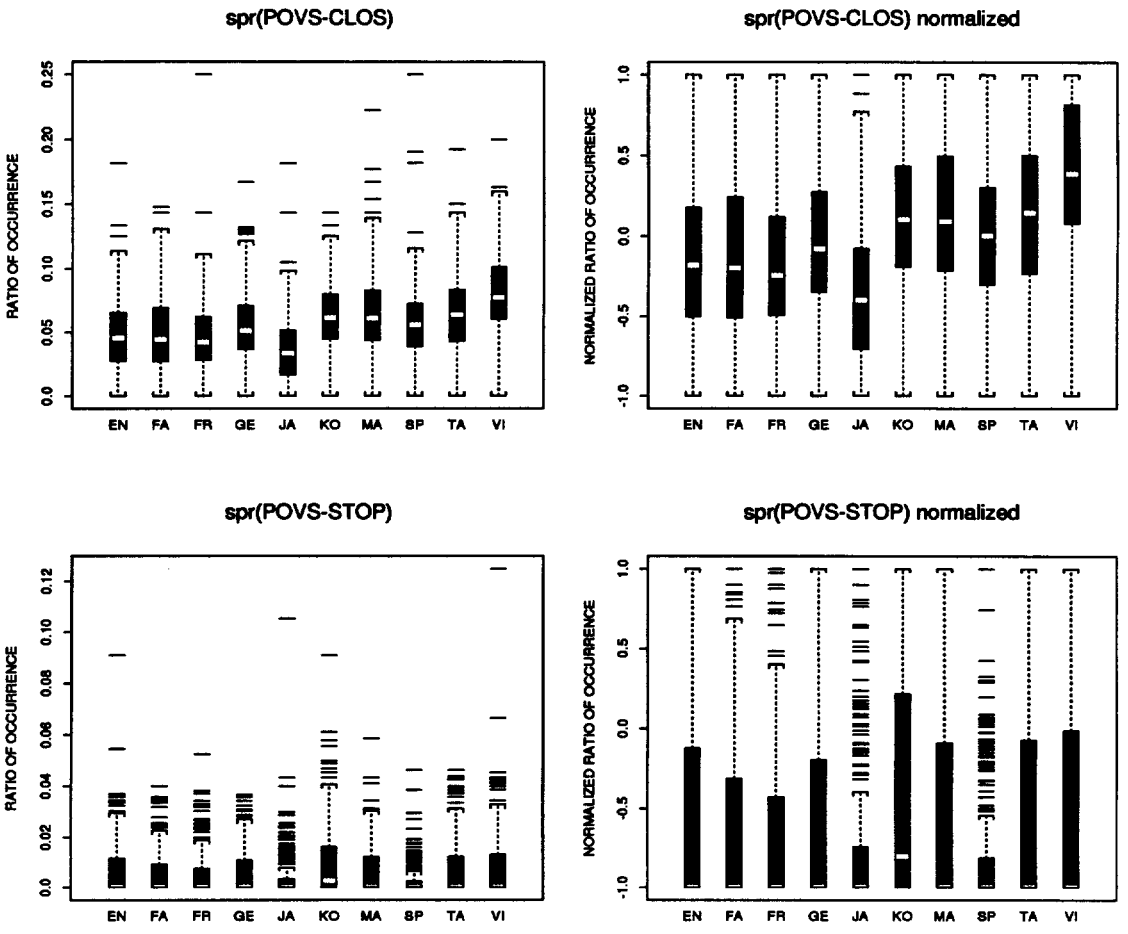

Figure C.56: Boxplots of SPR Features (continued)

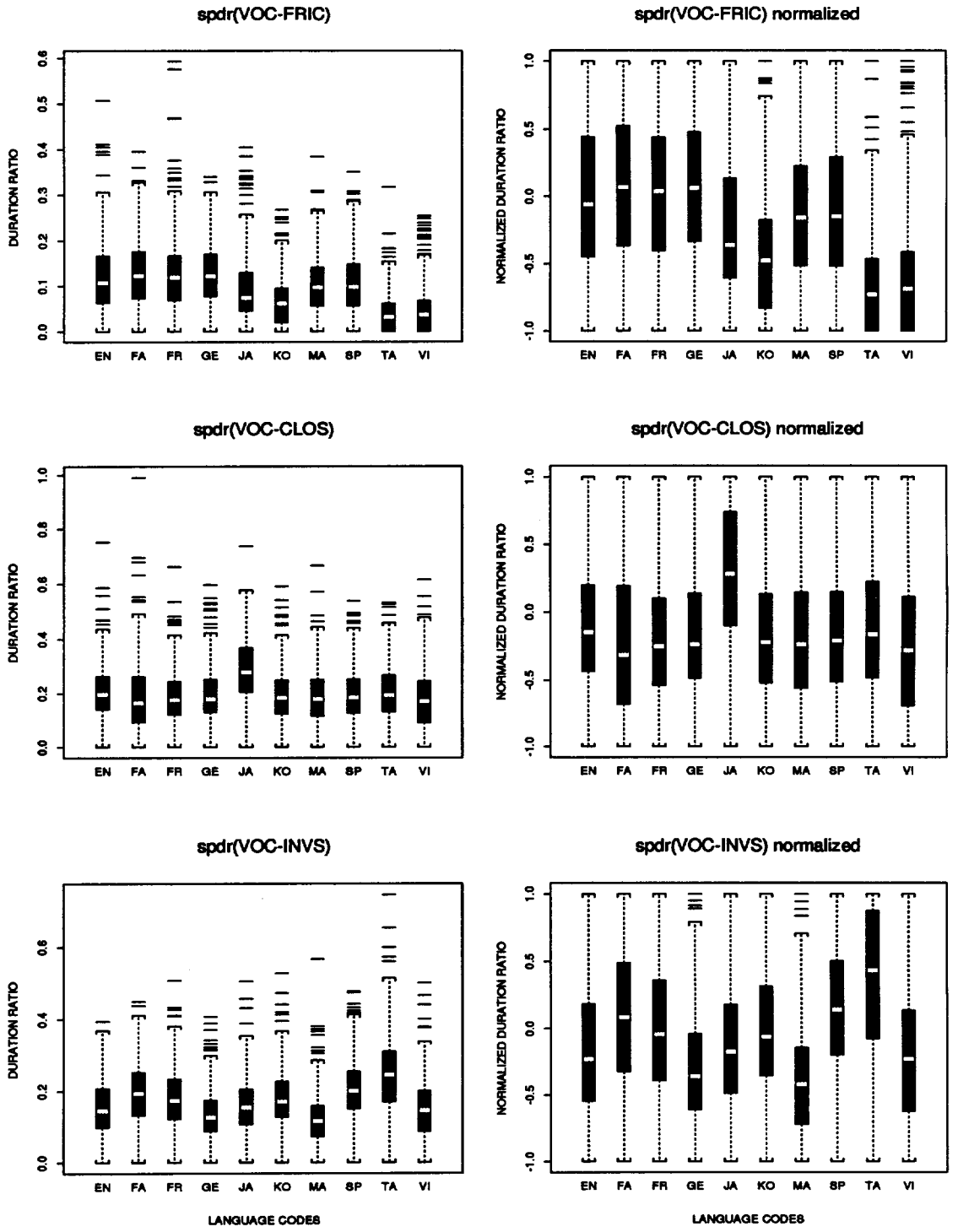

Figure C.57: Boxplots of Segment-pair Duration Ratio (SPDR) Features

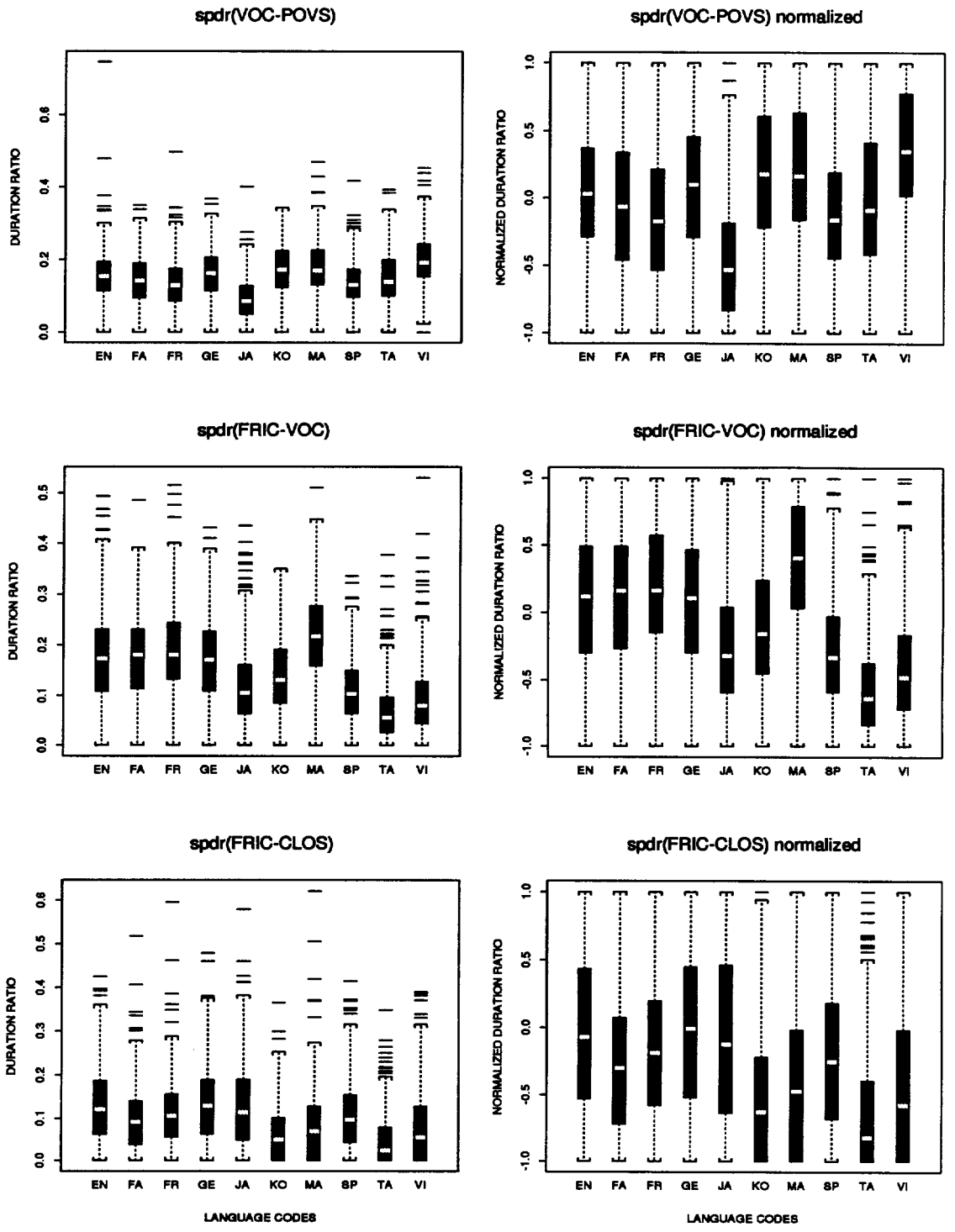

Figure C.58: Boxplots of SPDR Features (continued)

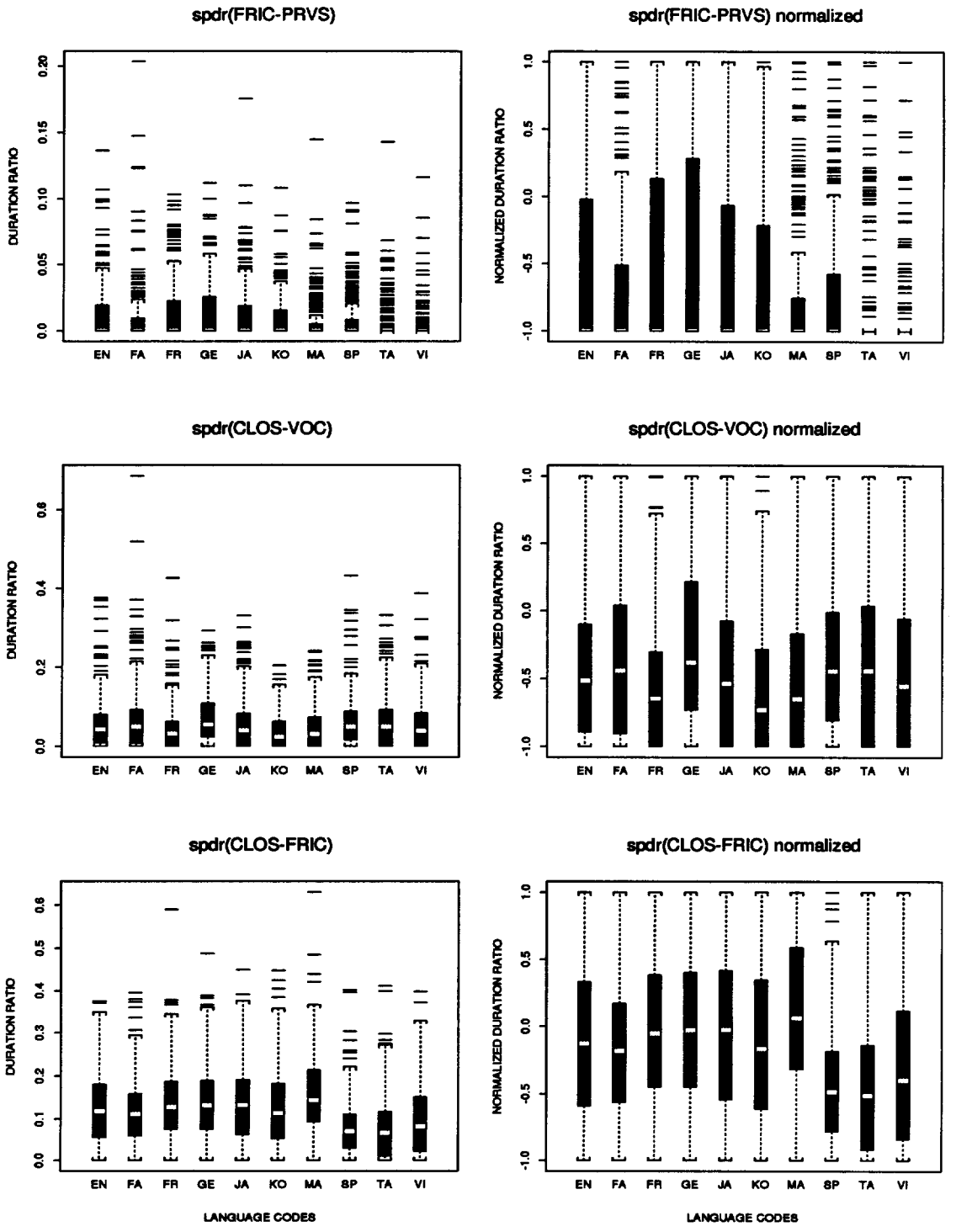

Figure C.59: Boxplots of SPDR Features (continued)

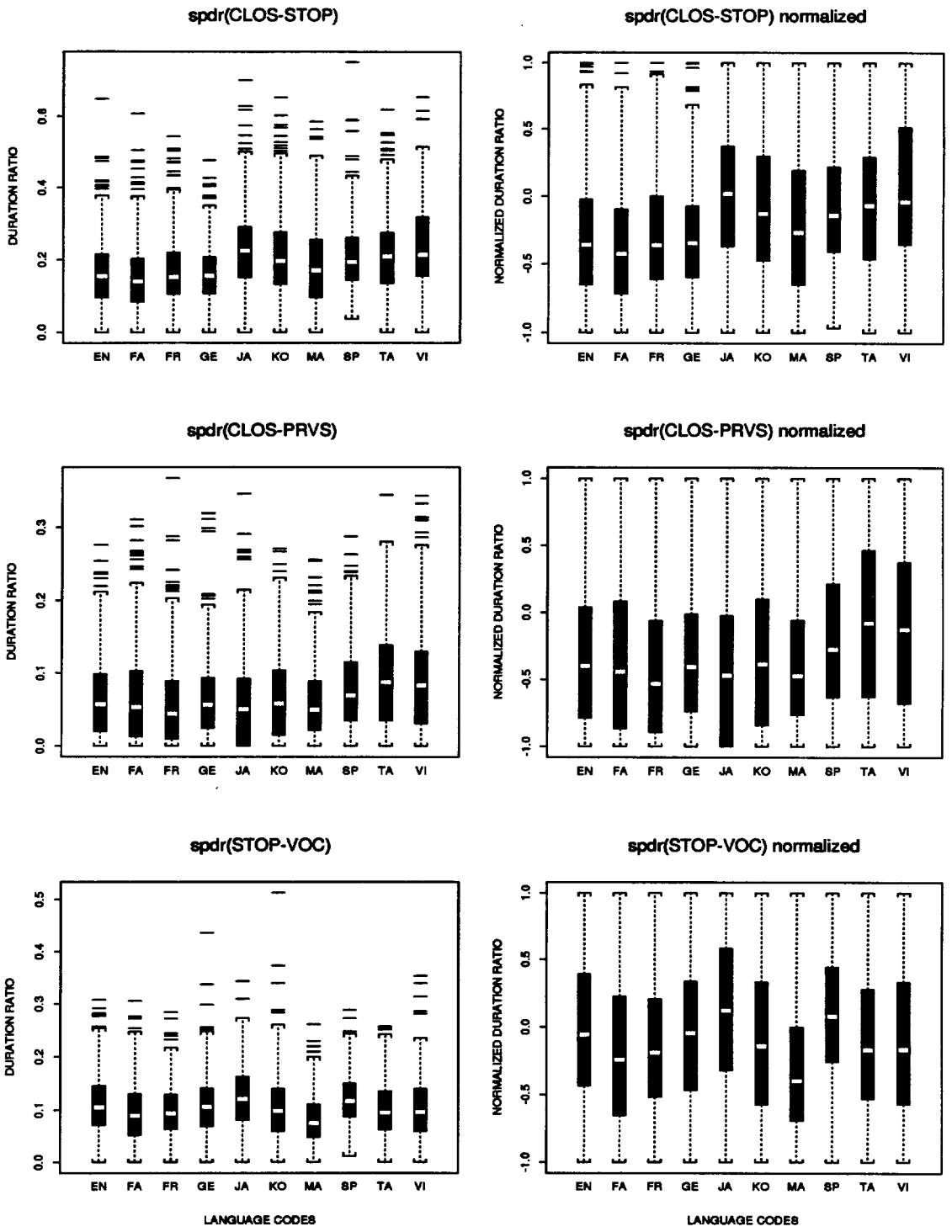

Figure C.60: Boxplots of SPDR Features (continued)

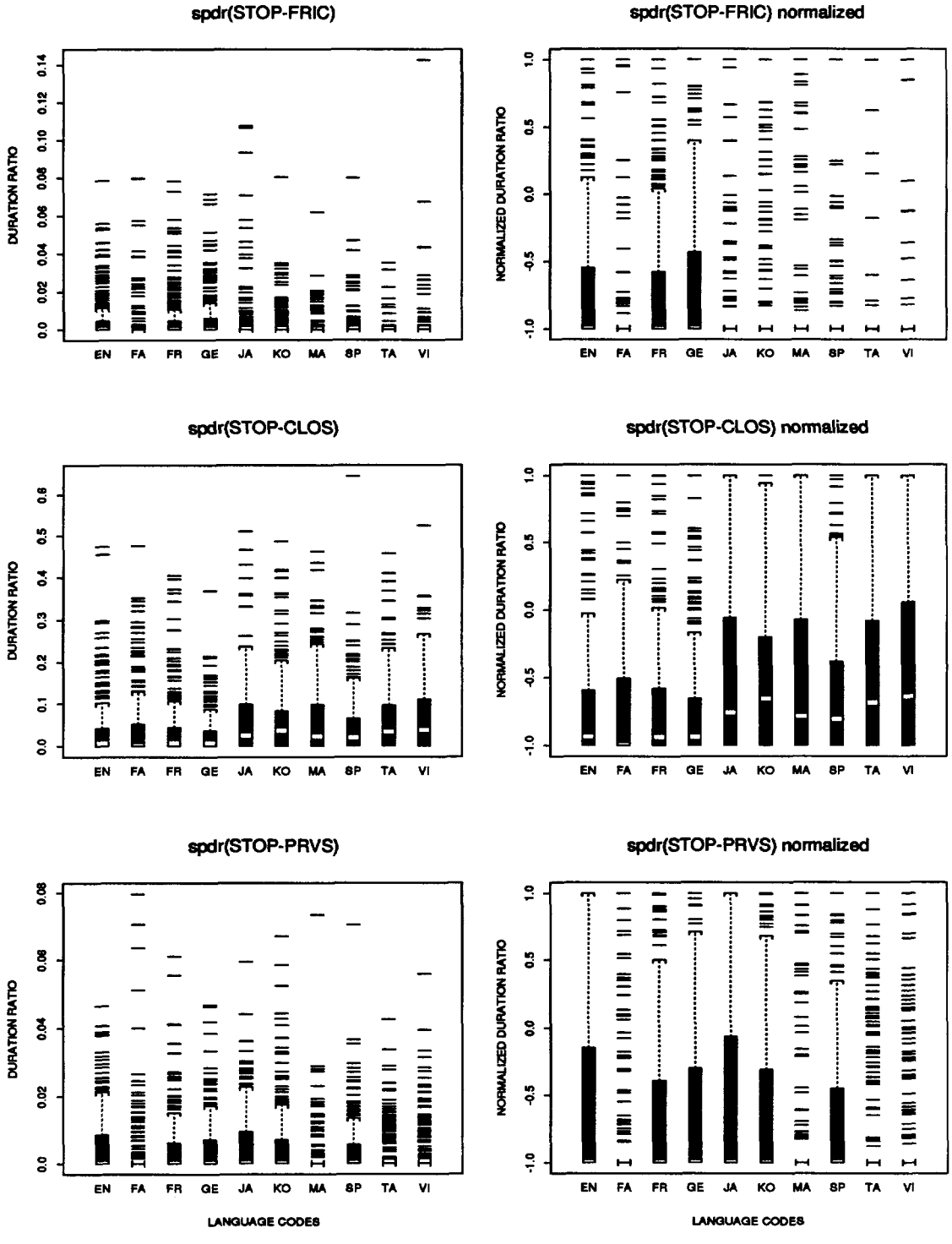

Figure C.61: Boxplots of SPDR Features (continued)

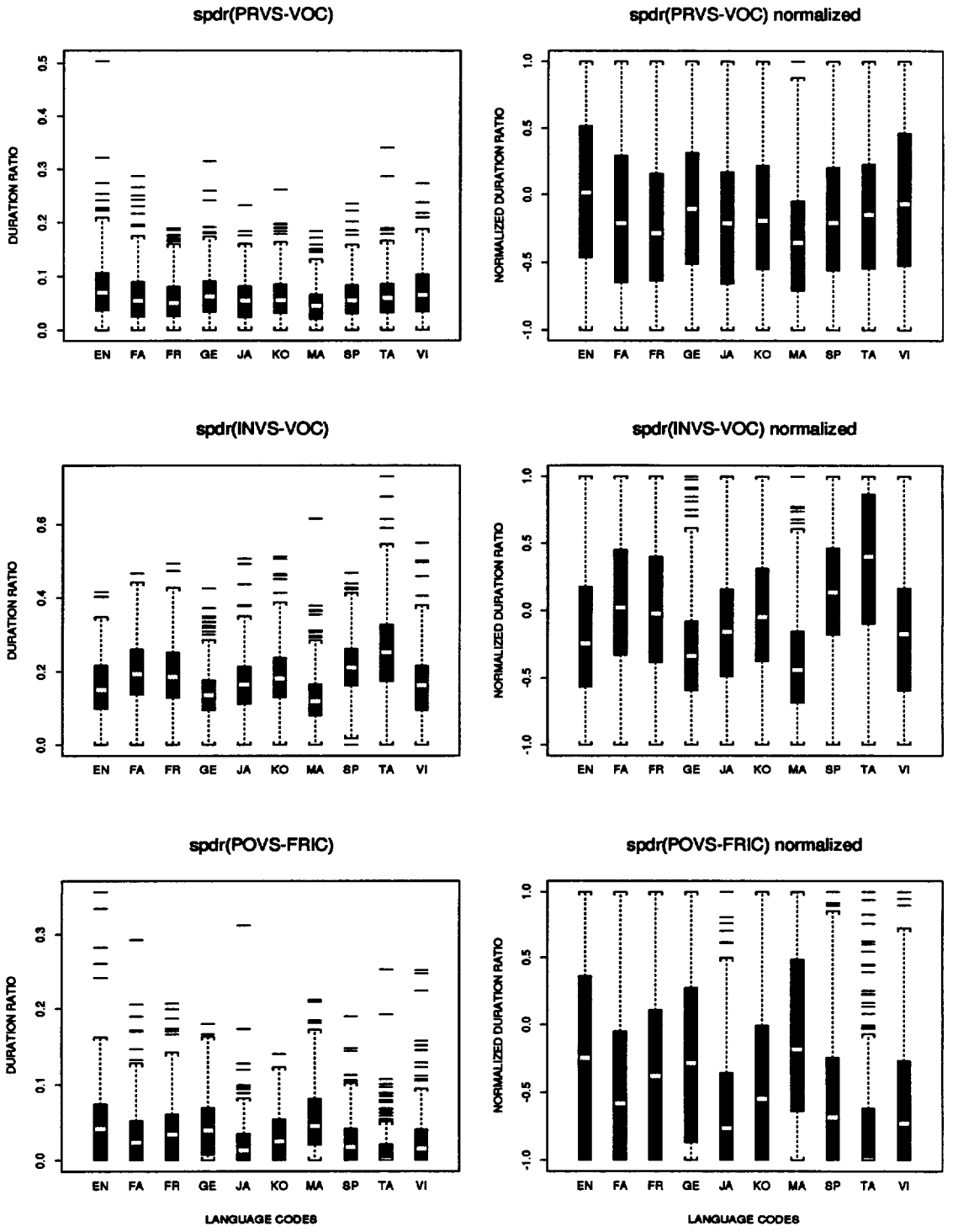

Figure C.62: Boxplots of SPDR Features (continued)

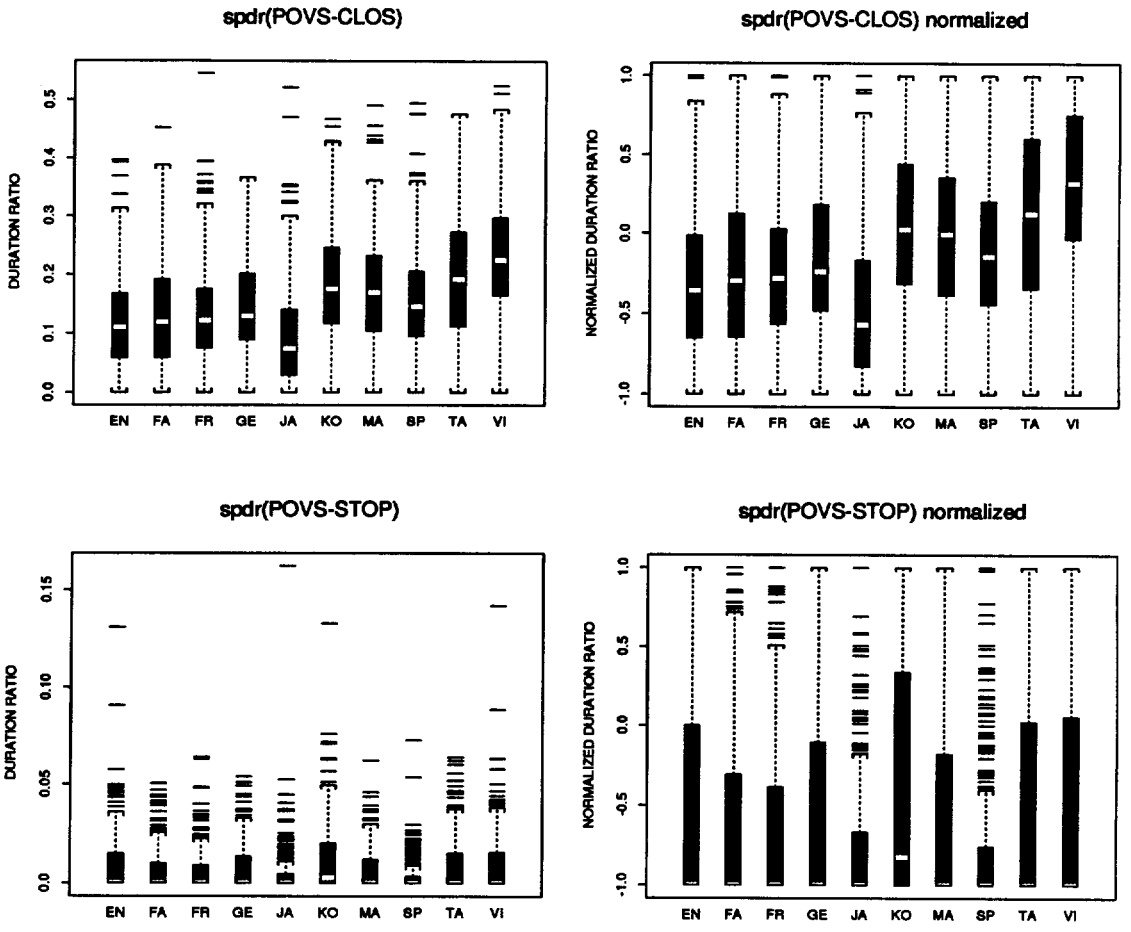

Figure C.63: Boxplots of SPDR Features (continued)

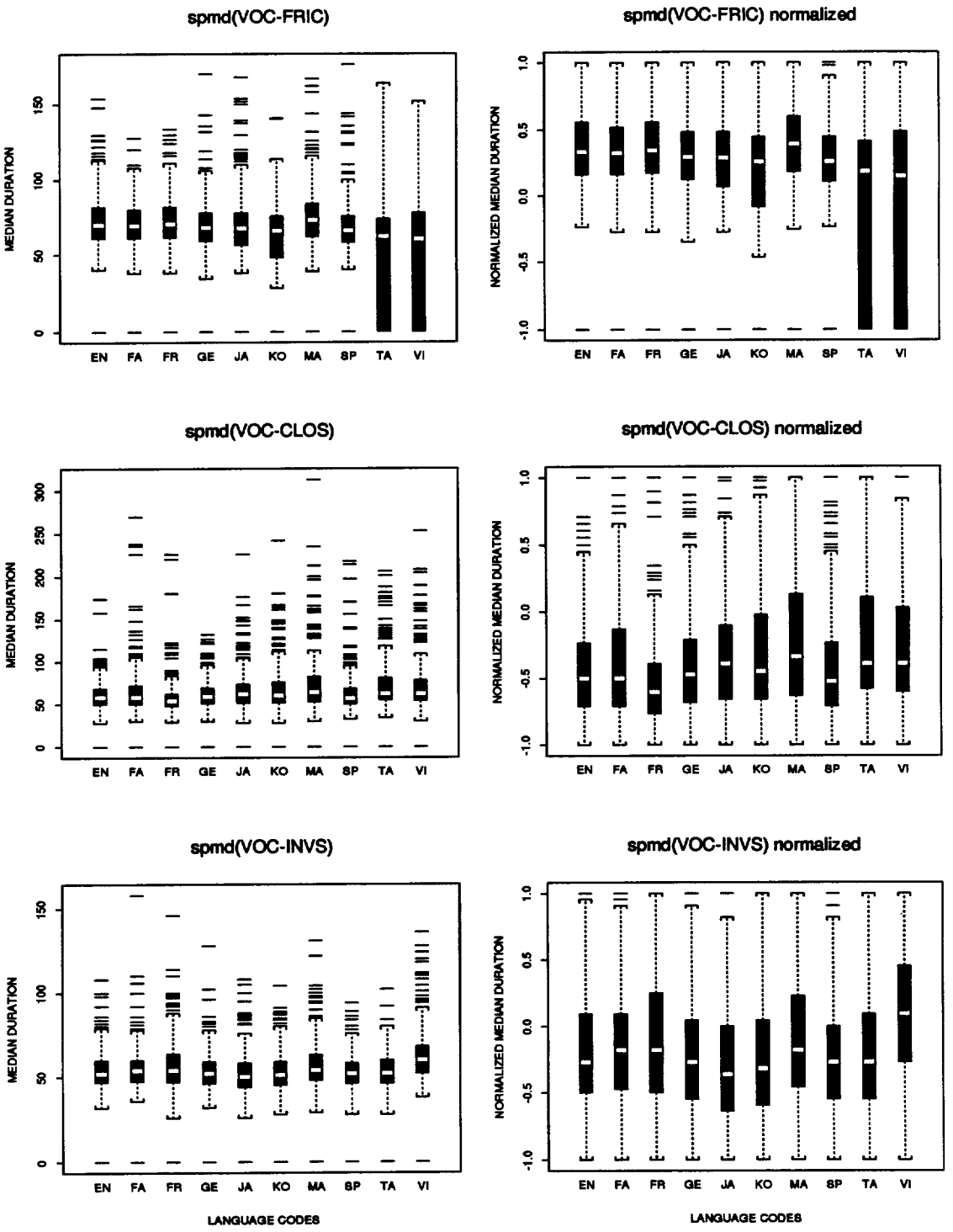

Figure C.64: Boxplots of Segment-pair Median Duration (SPMD) Features

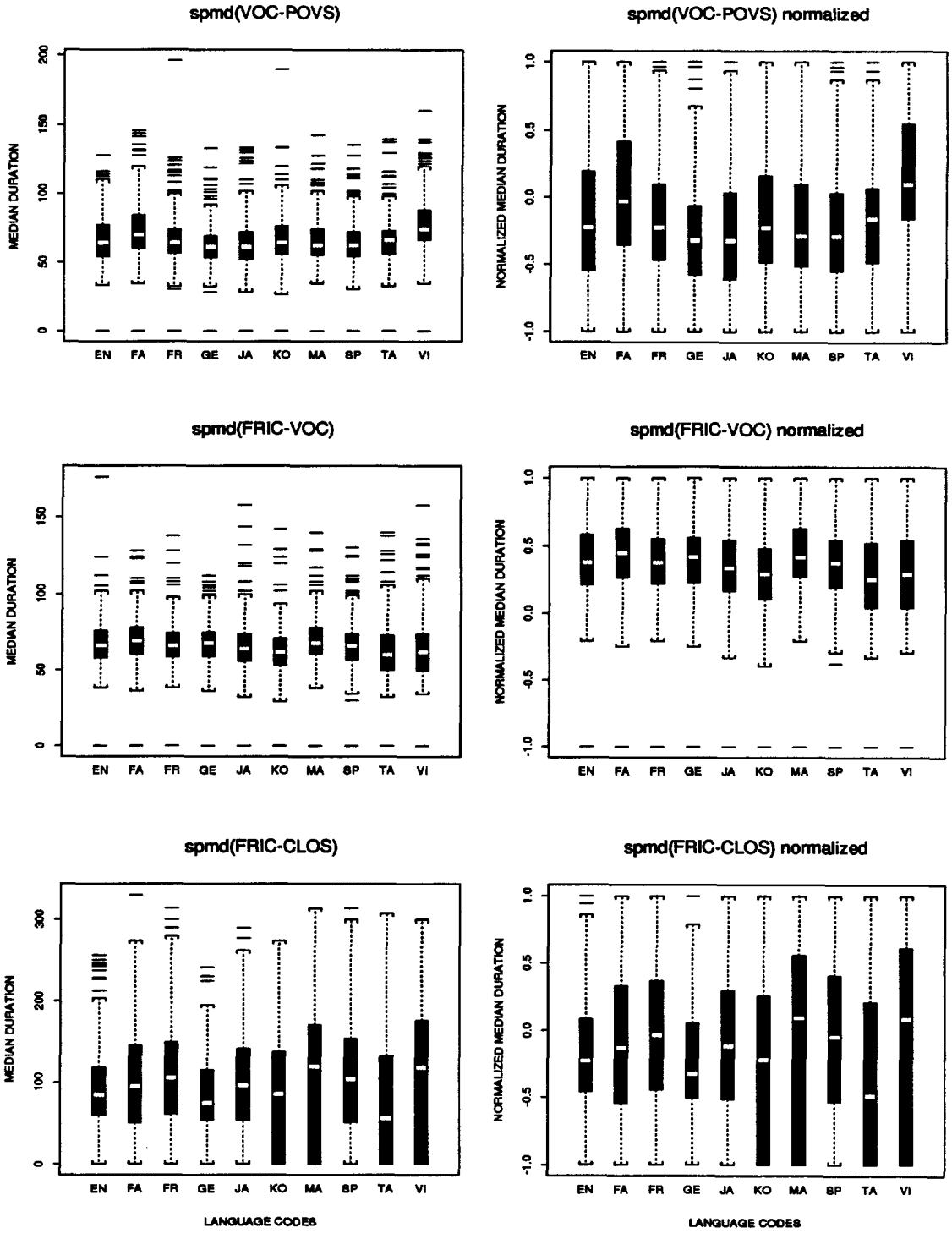

Figure C.65: Boxplots of SPMD Features (continued)

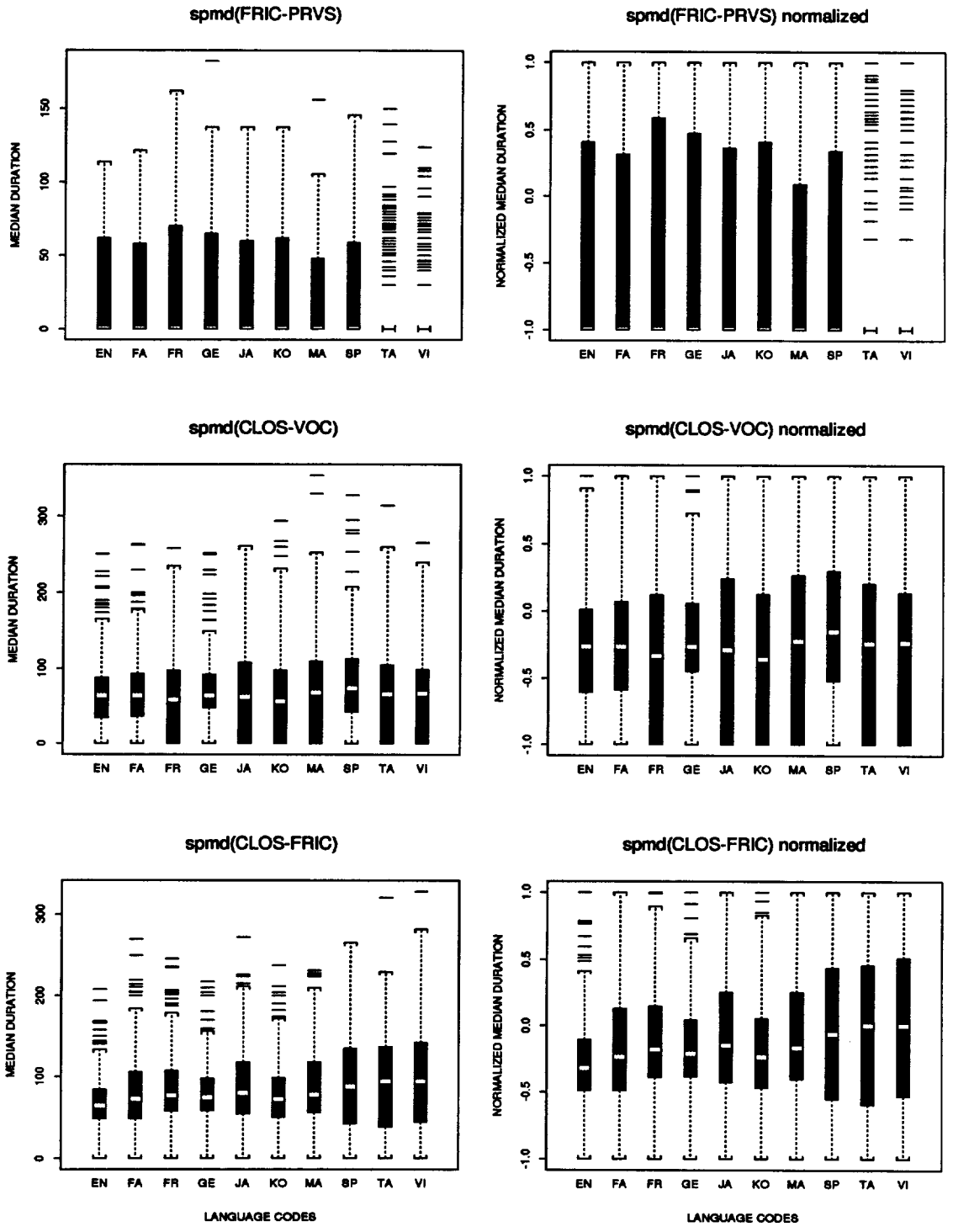

Figure C.66: Boxplots of SPMD Features (continued)

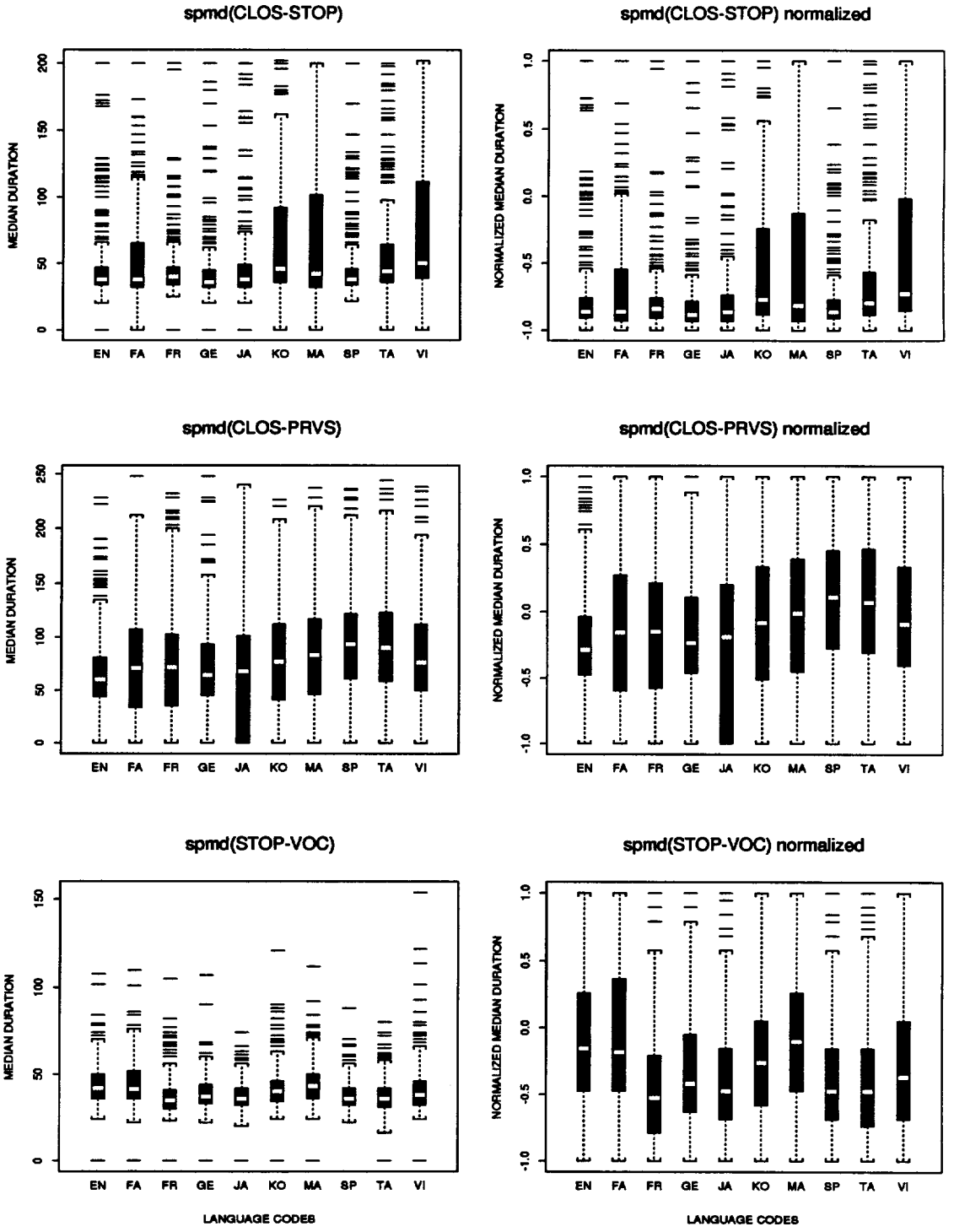

Figure C.67: Boxplots of SPMD Features (continued)

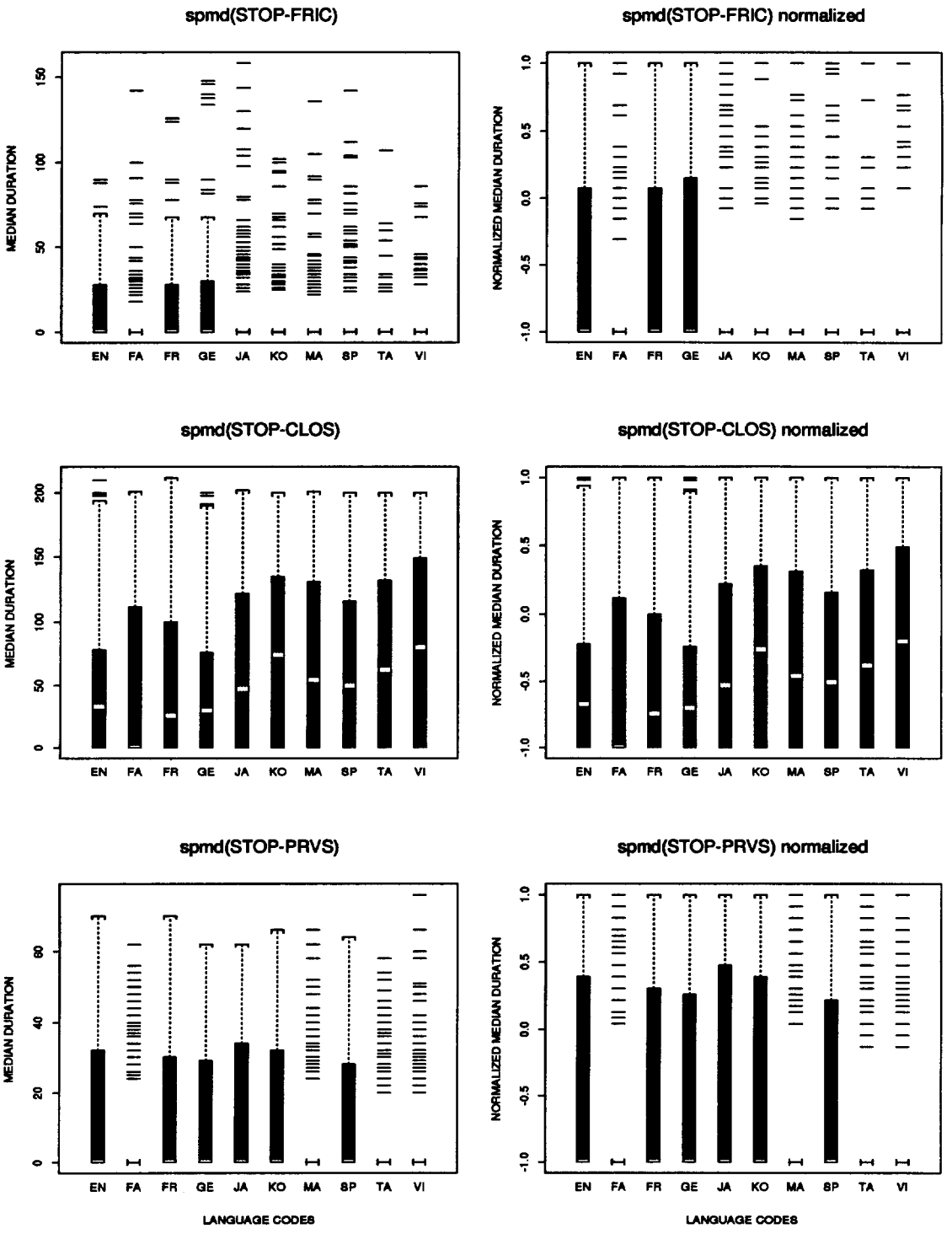

Figure C.68: Boxplots of SPMD Features (continued)

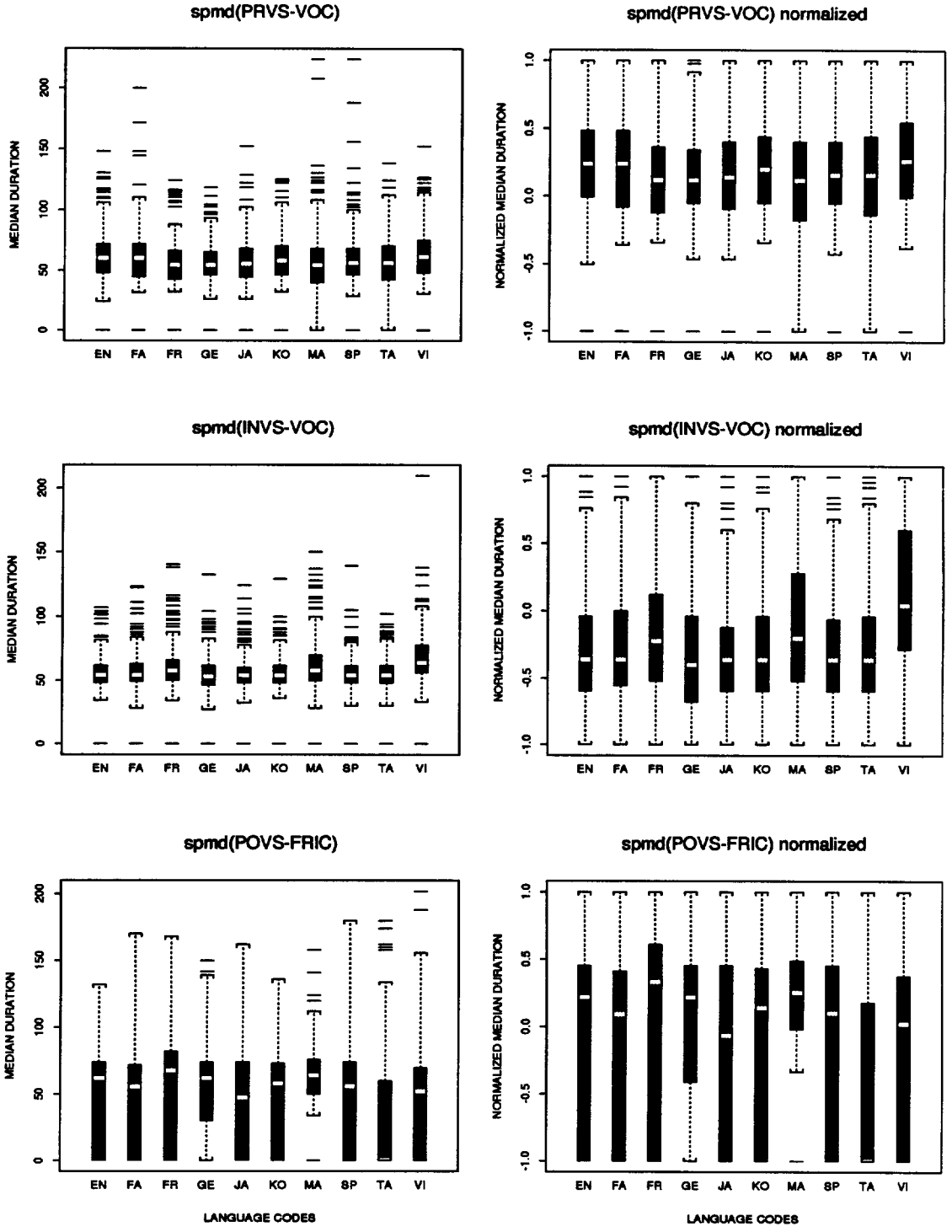

Figure C.69: Boxplots of SPMD Features (continued)

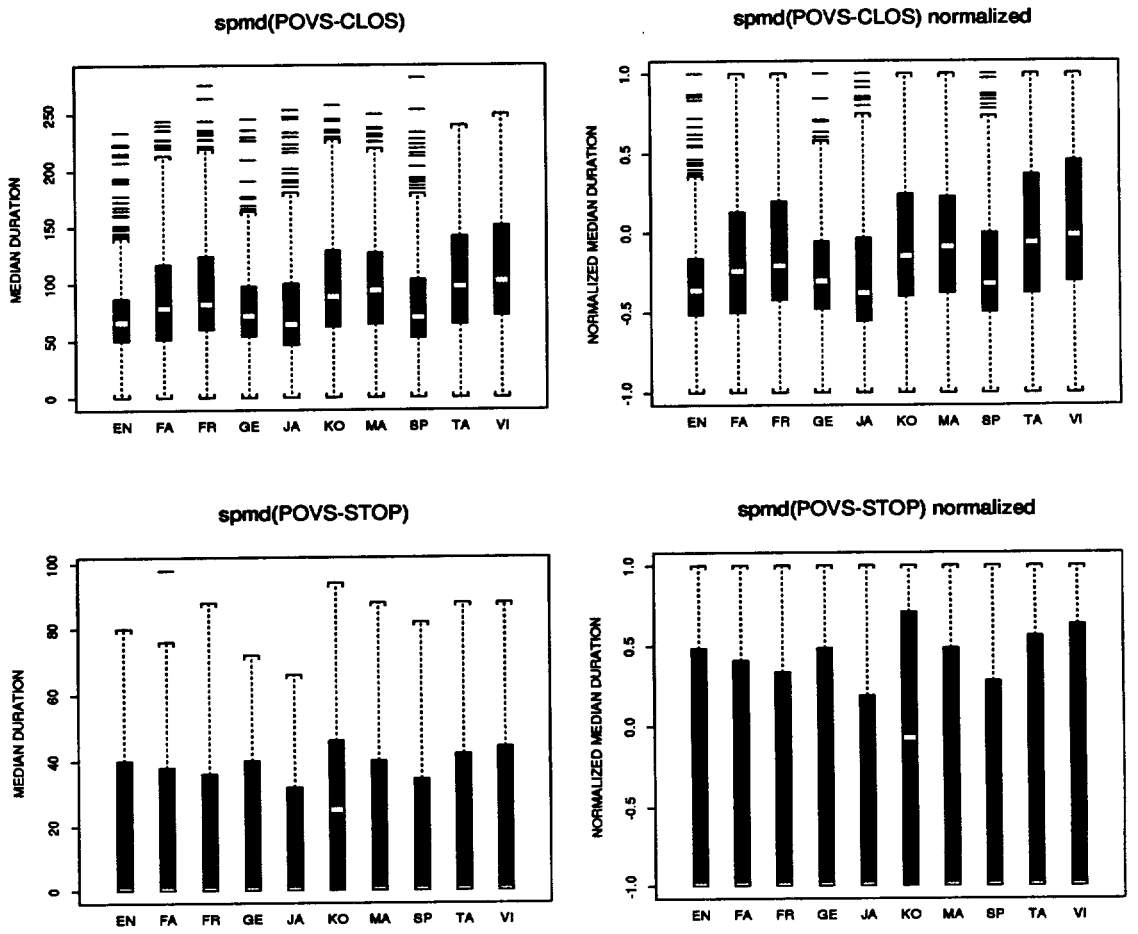

Figure C.70: Boxplots of SPMD Features (continued)

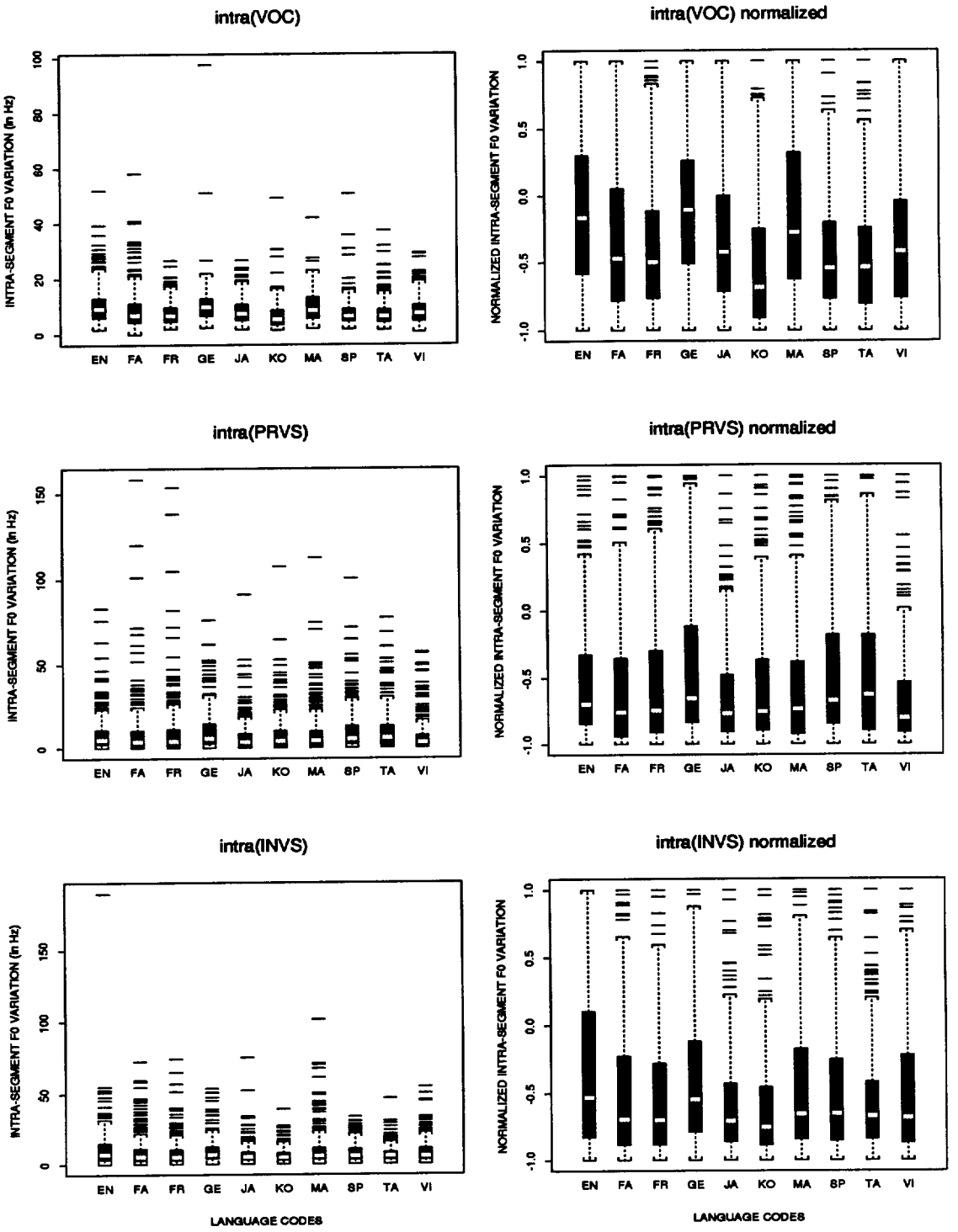

Figure C.71: Boxplots of Intra-segment Variation in Pitch

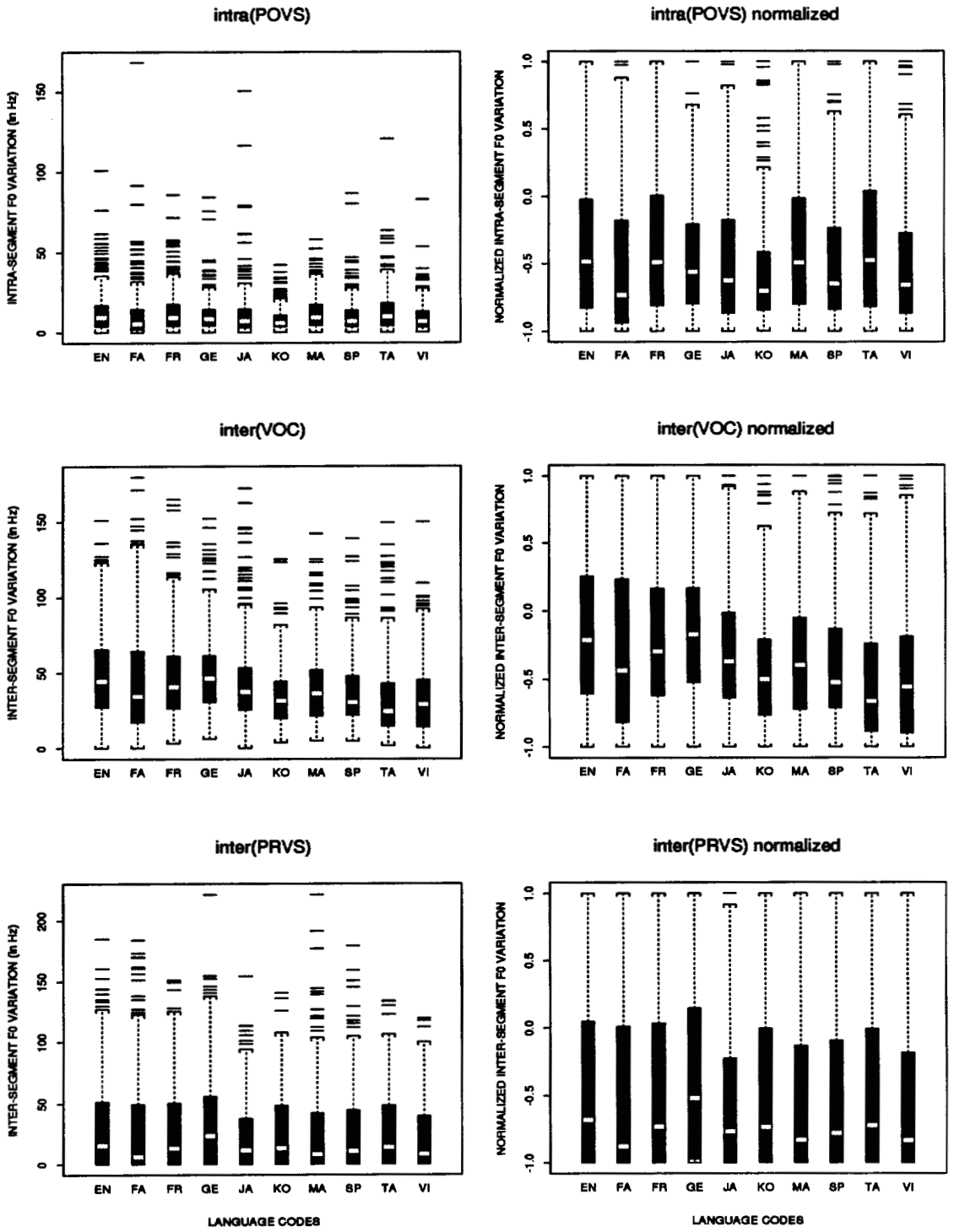

Figure C.72: Boxplots of Intra-segment and Inter-segment Variation in Pitch

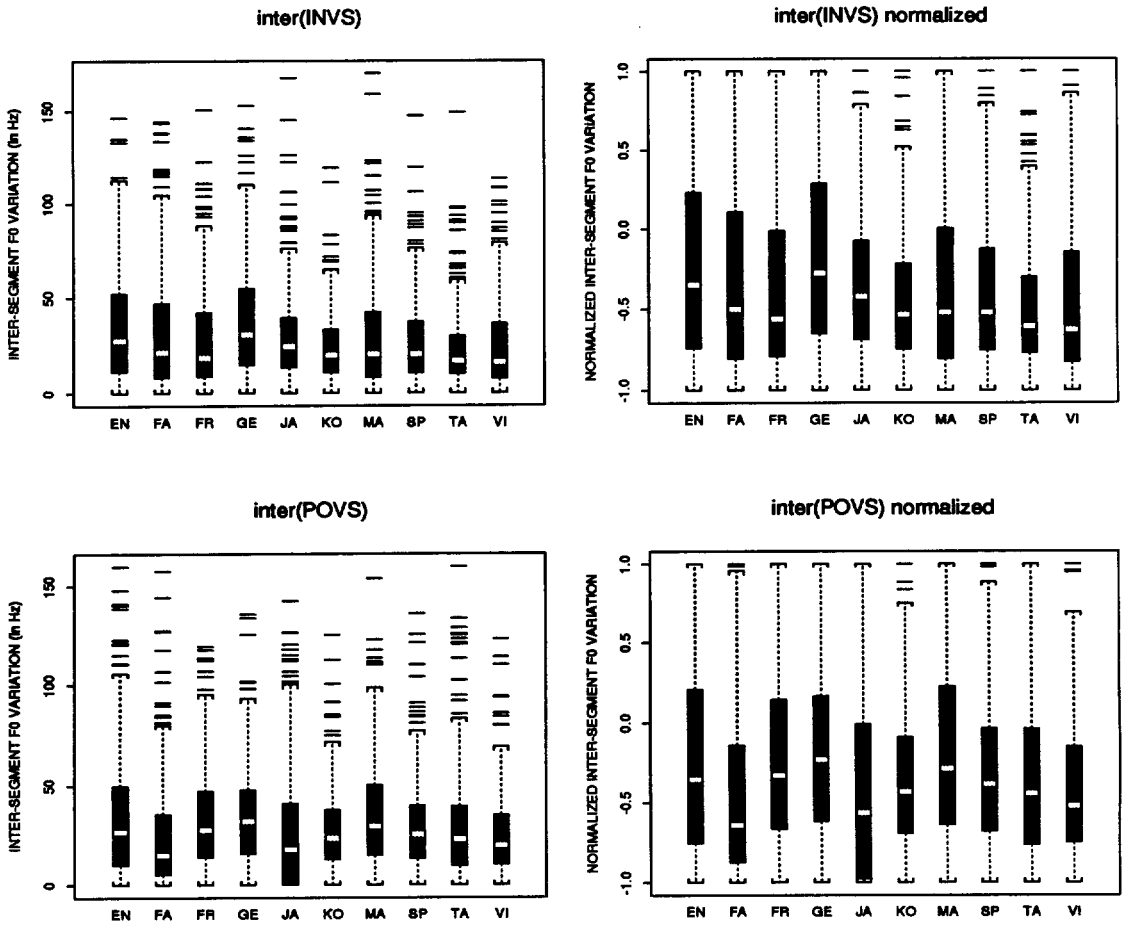

Figure C.73: Boxplots of Inter-segment Variation in Pitch (continued)

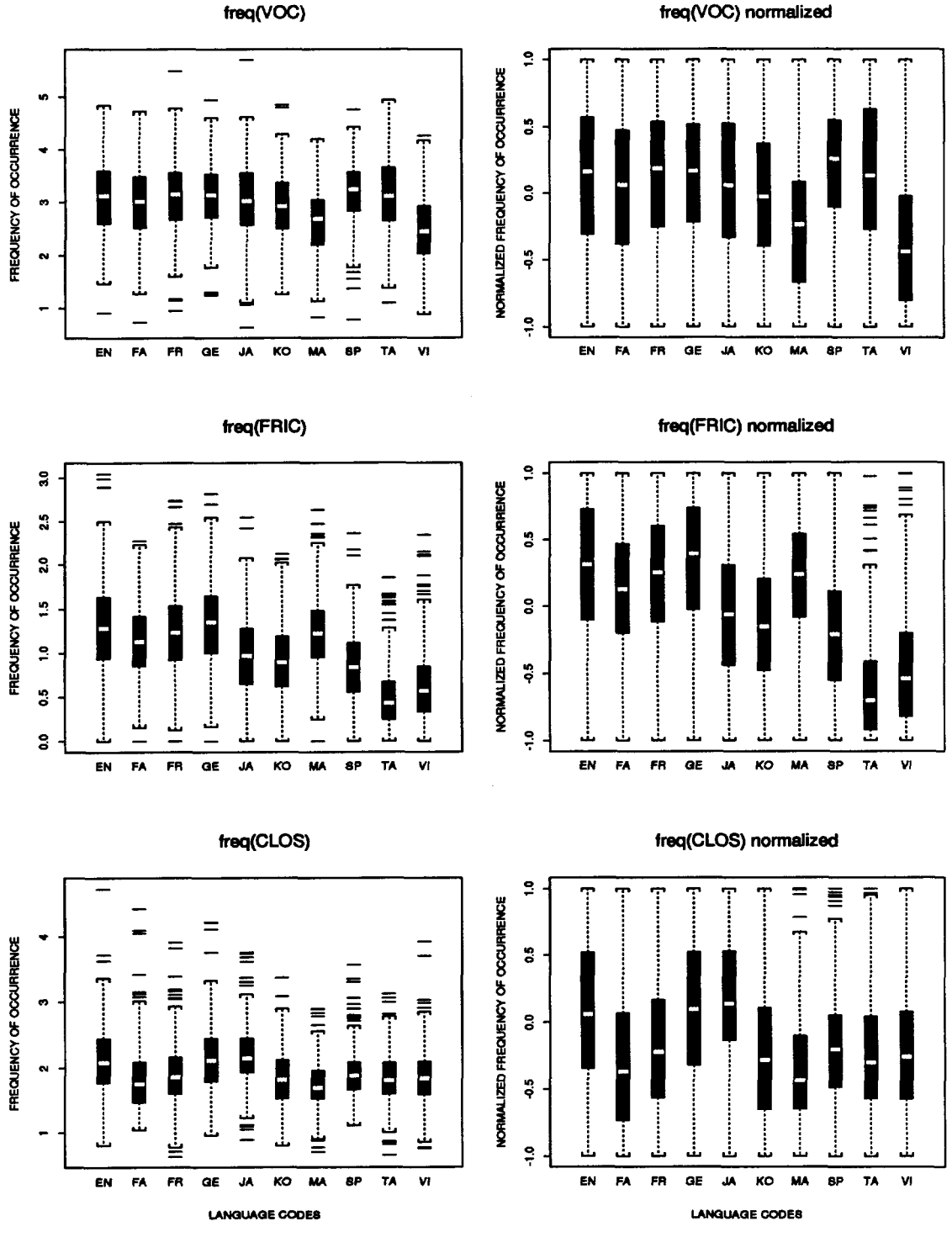

Figure C.74: Frequency of Occurrence of VOC, FRIC and CLOS

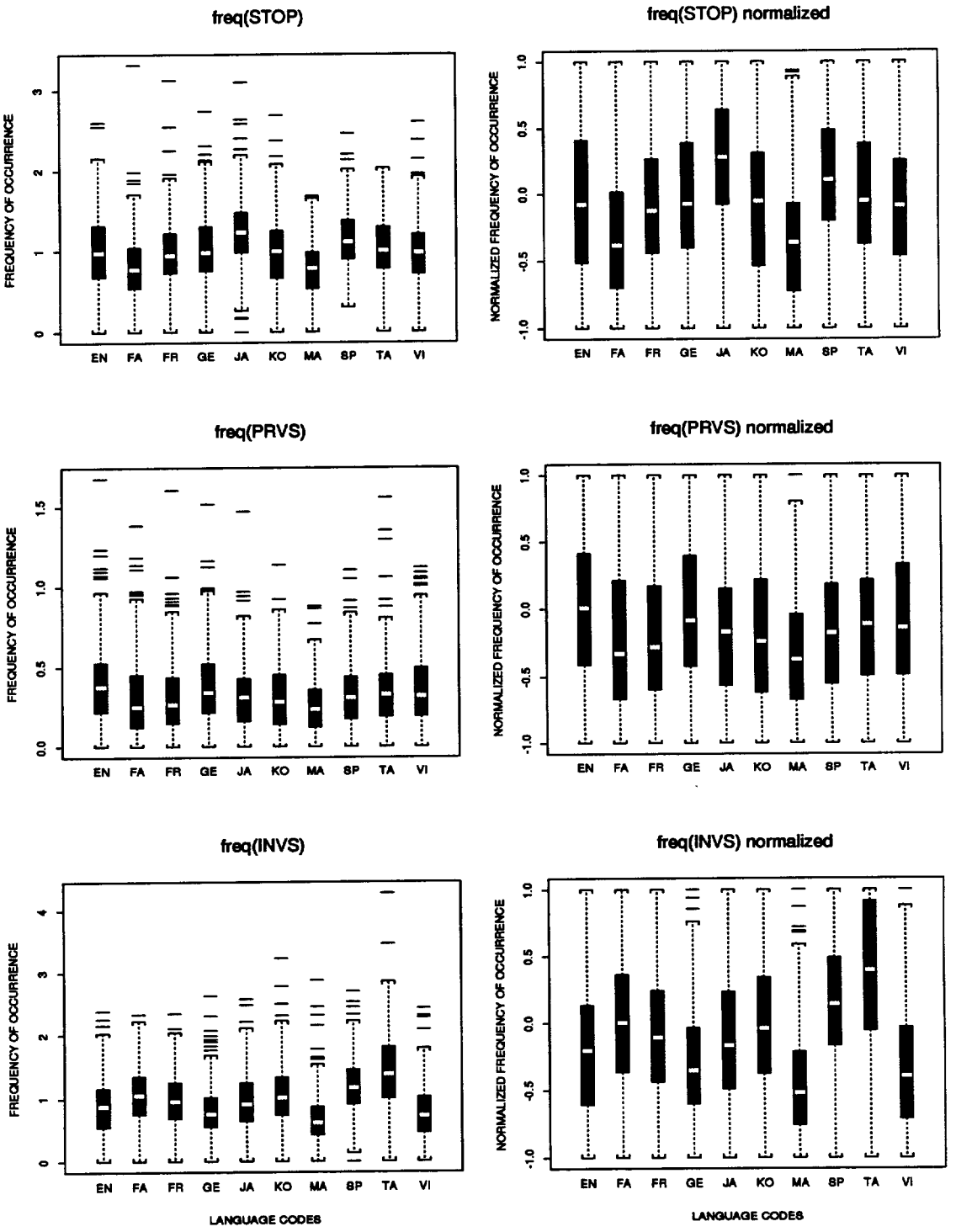

Figure C.75: Frequency of Occurrence of STOP, PRVS and INVS

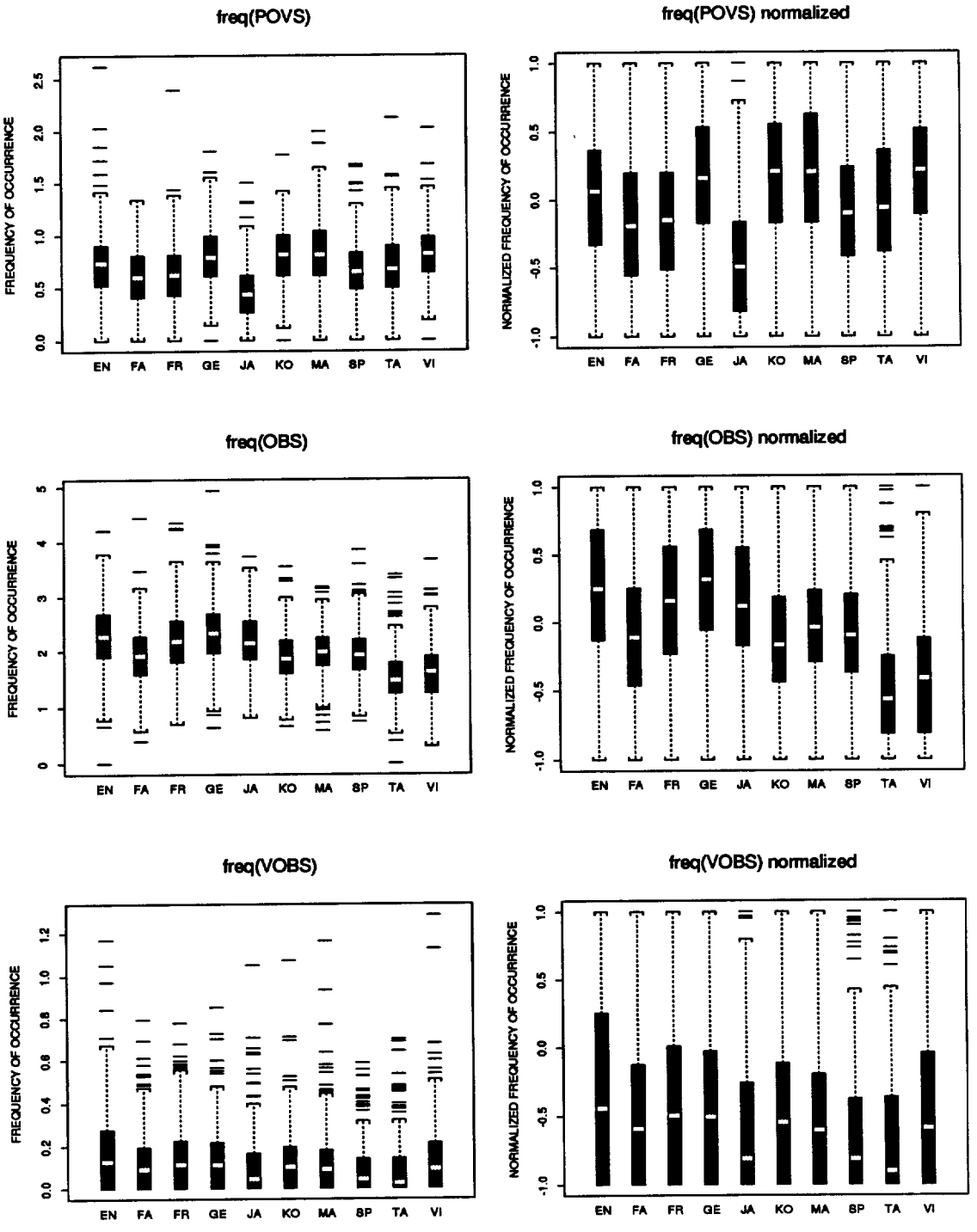

Figure C.76: Frequency of Occurrence of POVS, OBS and VOBS

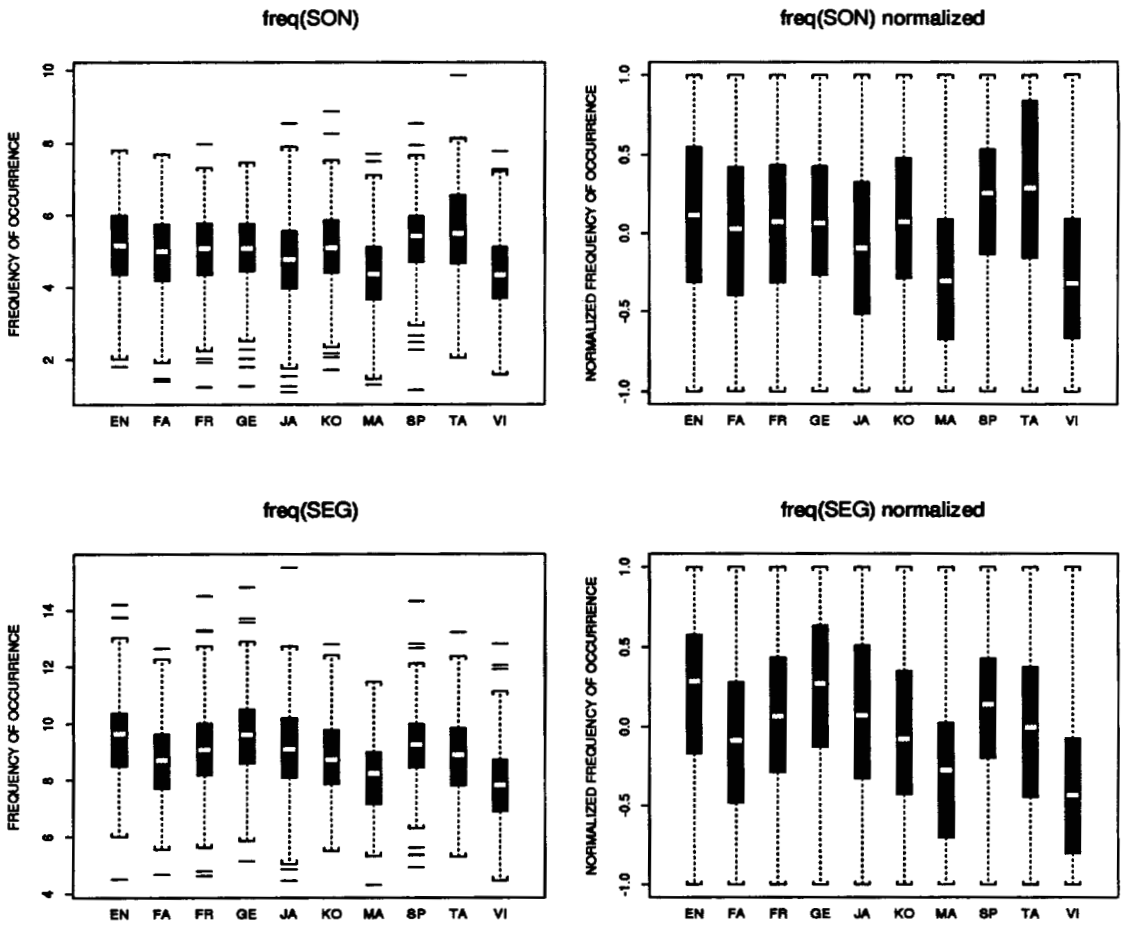

Figure C.77: Frequency of Occurrence of SON and SEG

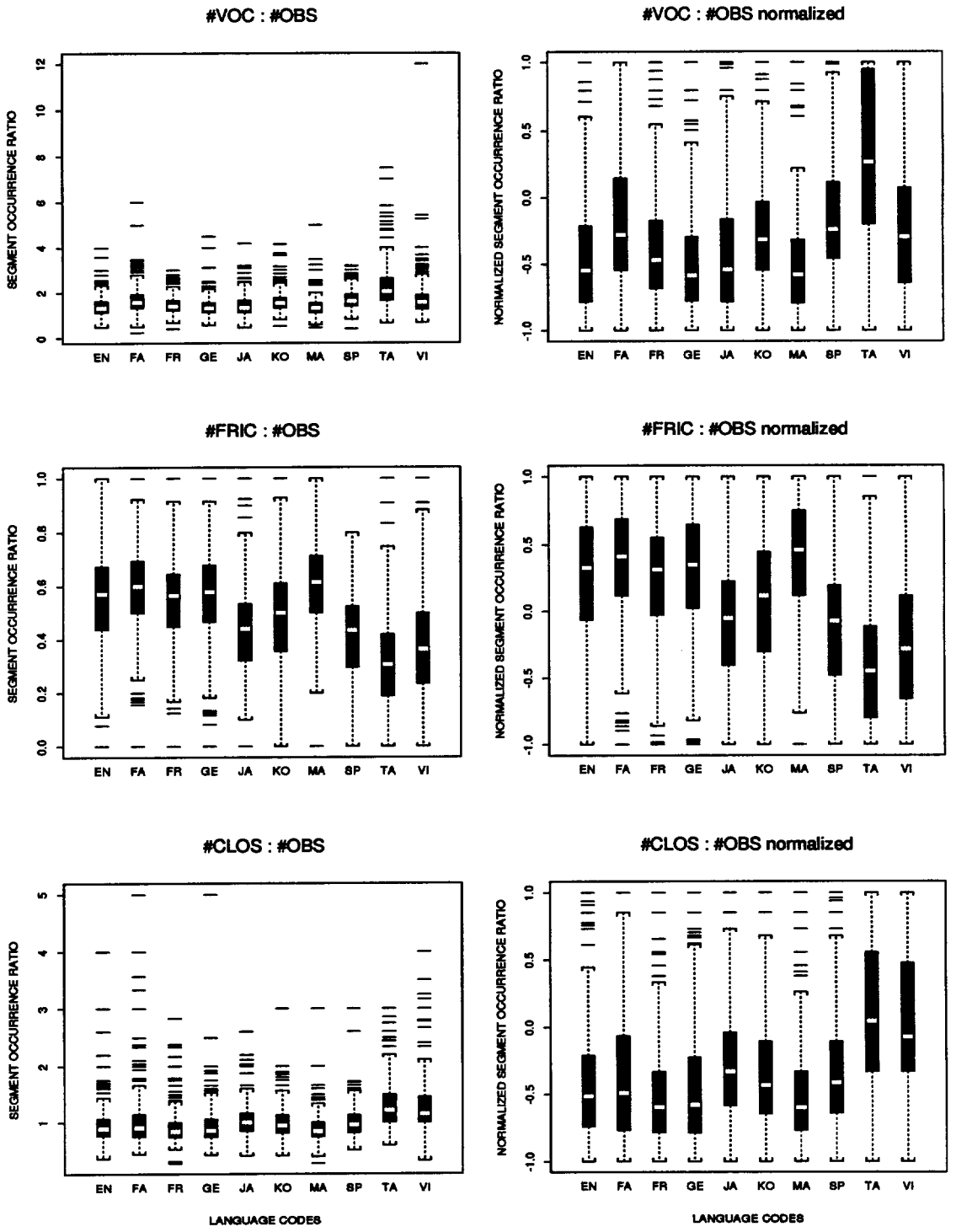

Figure C.78: Boxplots of Segment Occurrence Ratios

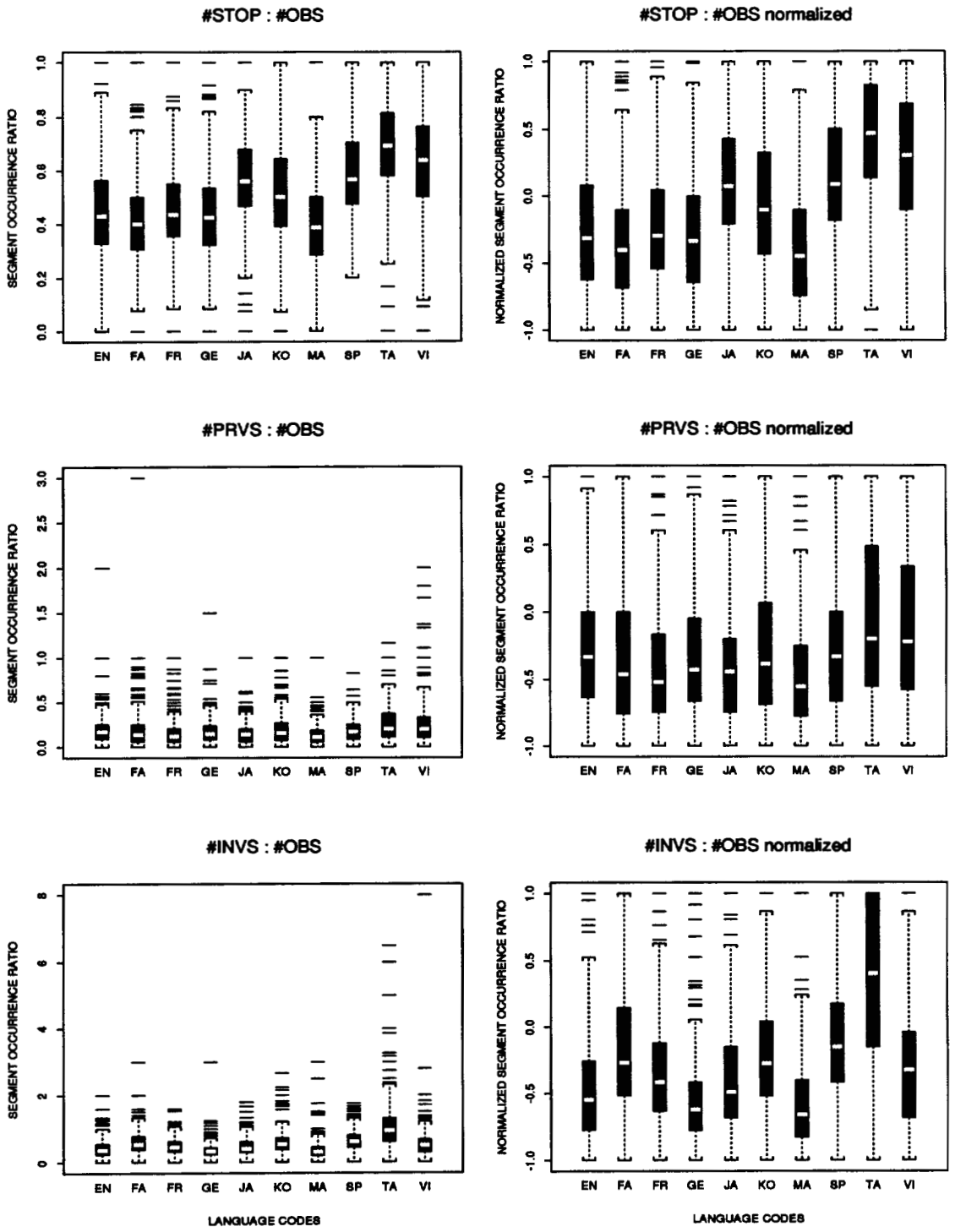

Figure C.79: Boxplots of Segment Occurrence Ratios (continued)

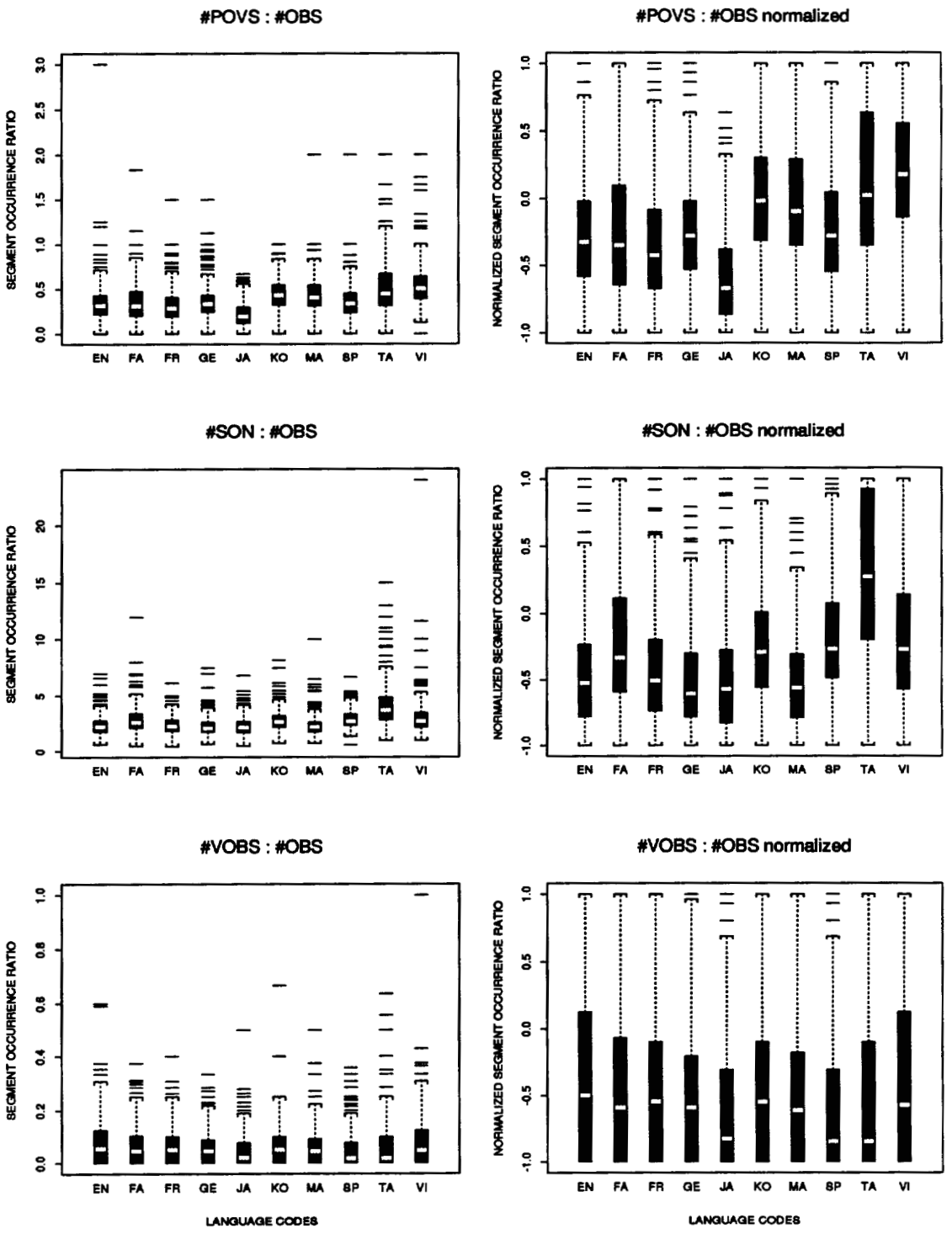

Figure C.80: Boxplots of Segment Occurrence Ratios (continued)

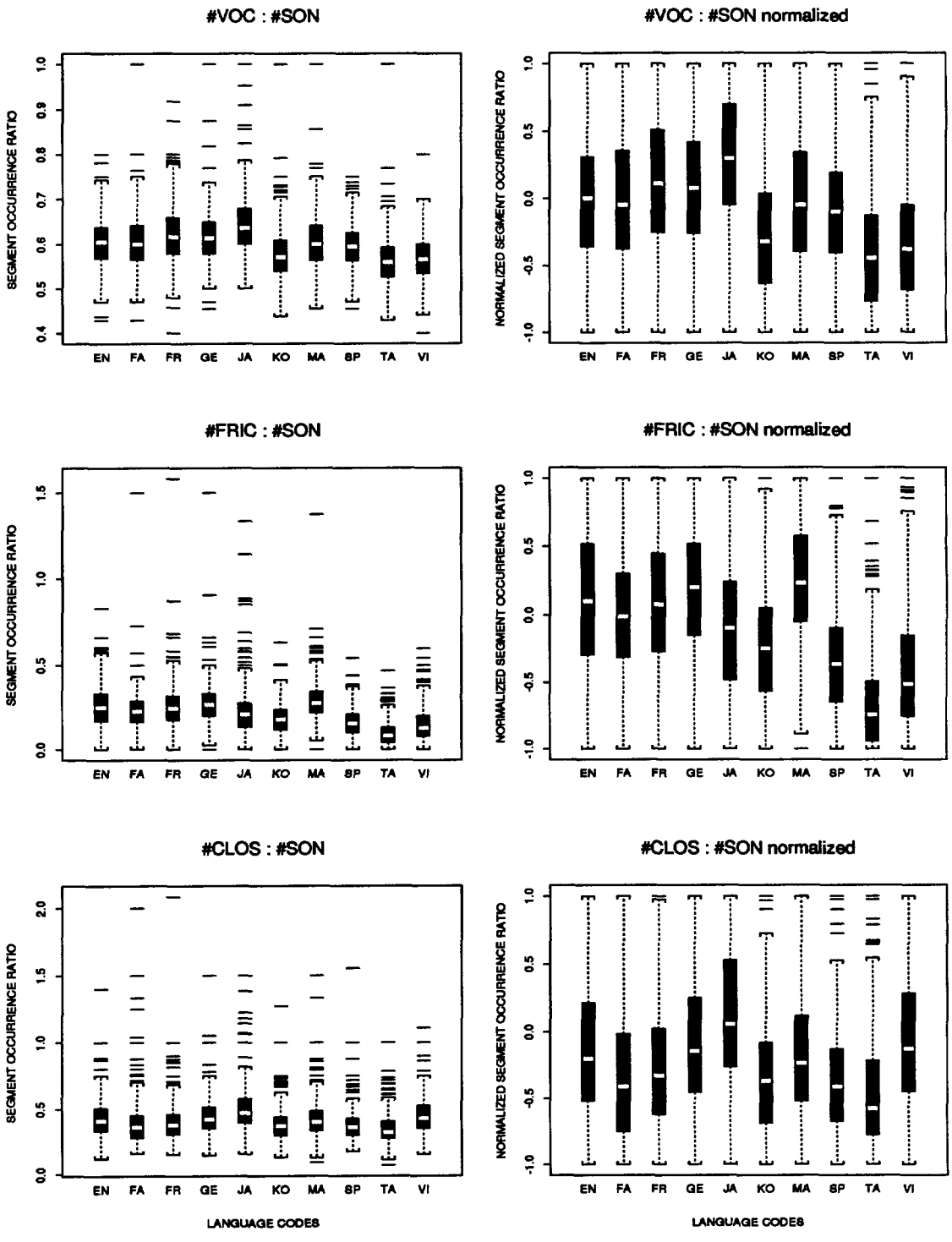

Figure C.81: Boxplots of Segment Occurrence Ratios (continued)

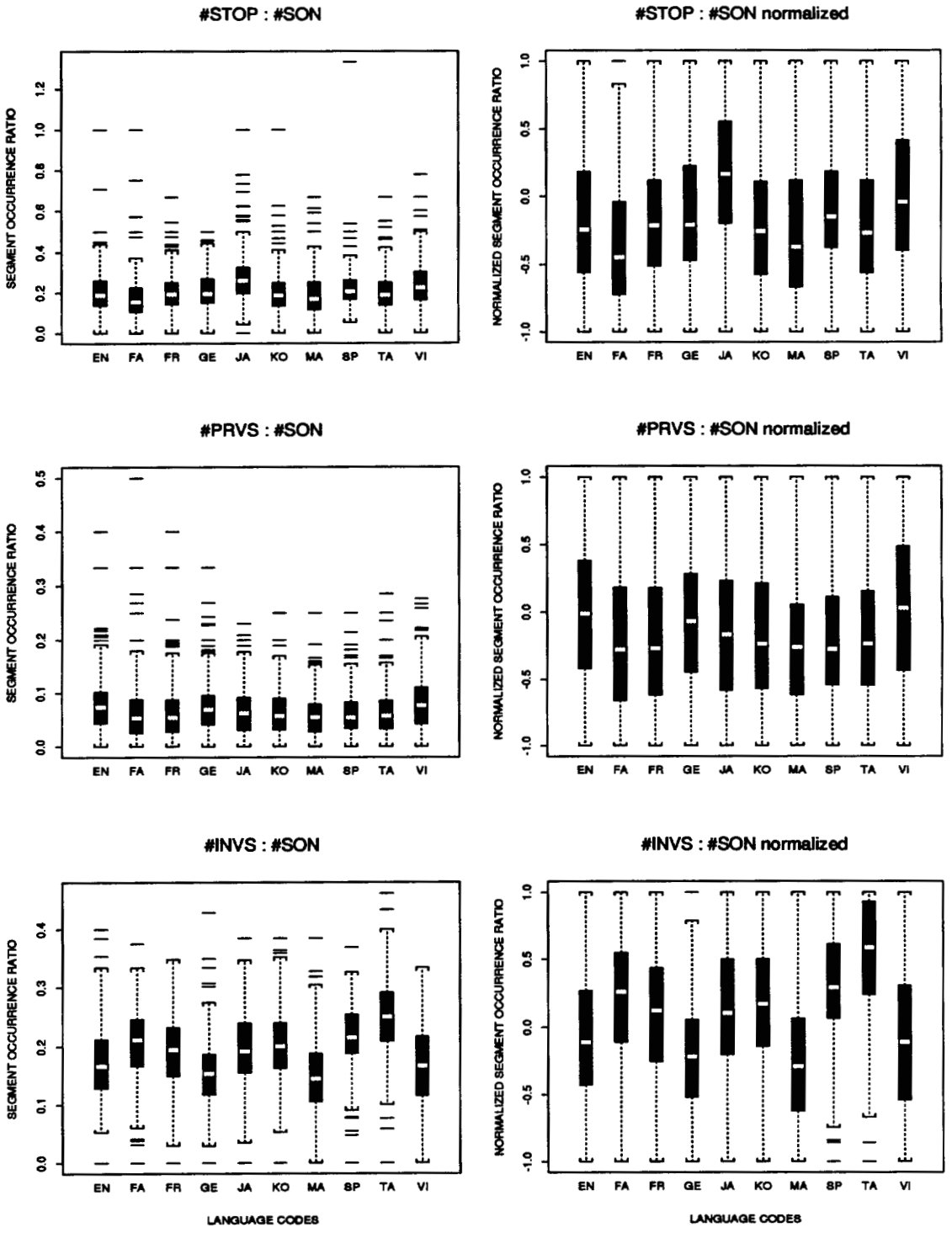

Figure C.82: Boxplots of Segment Occurrence Ratios (continued)

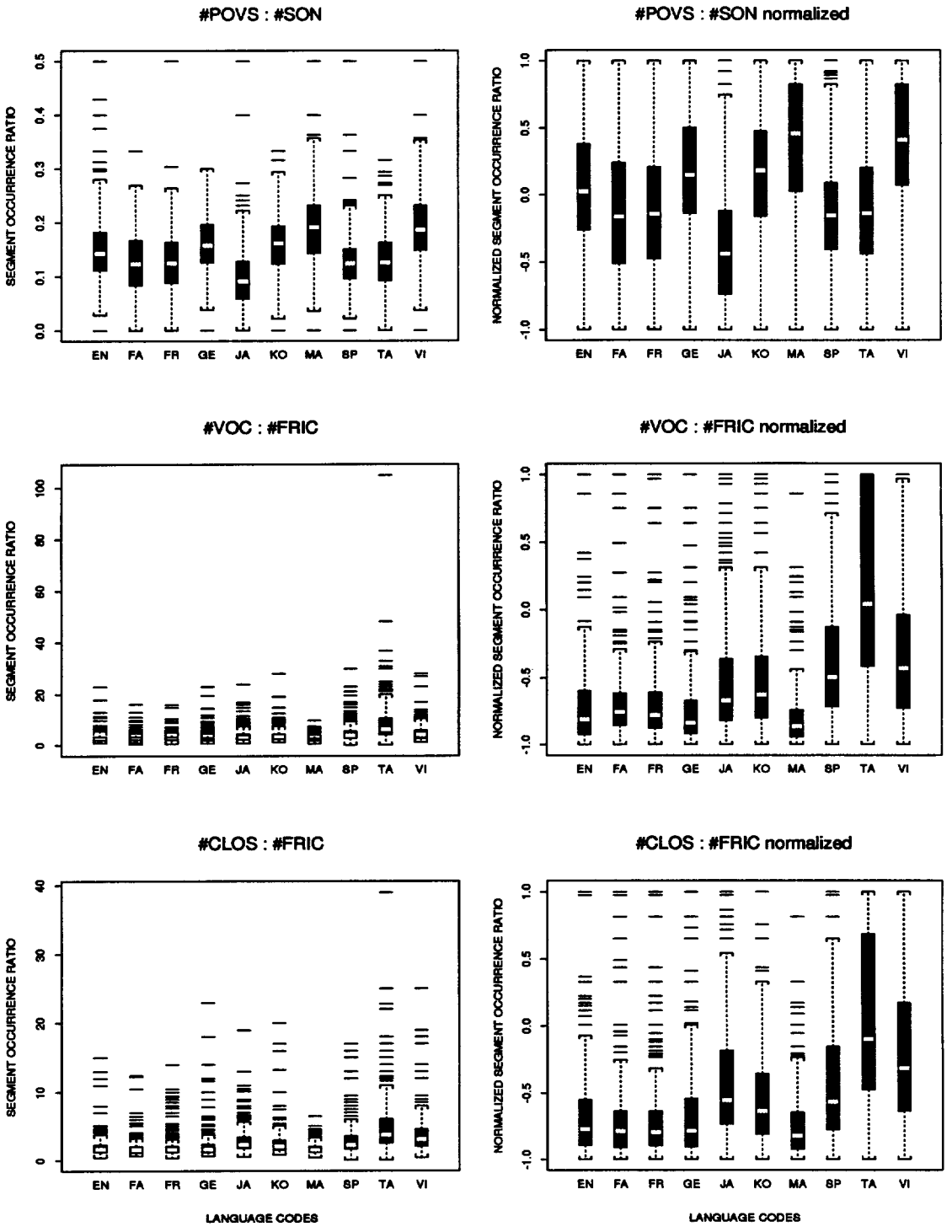

Figure C.83: Boxplots of Segment Occurrence Ratios (continued)

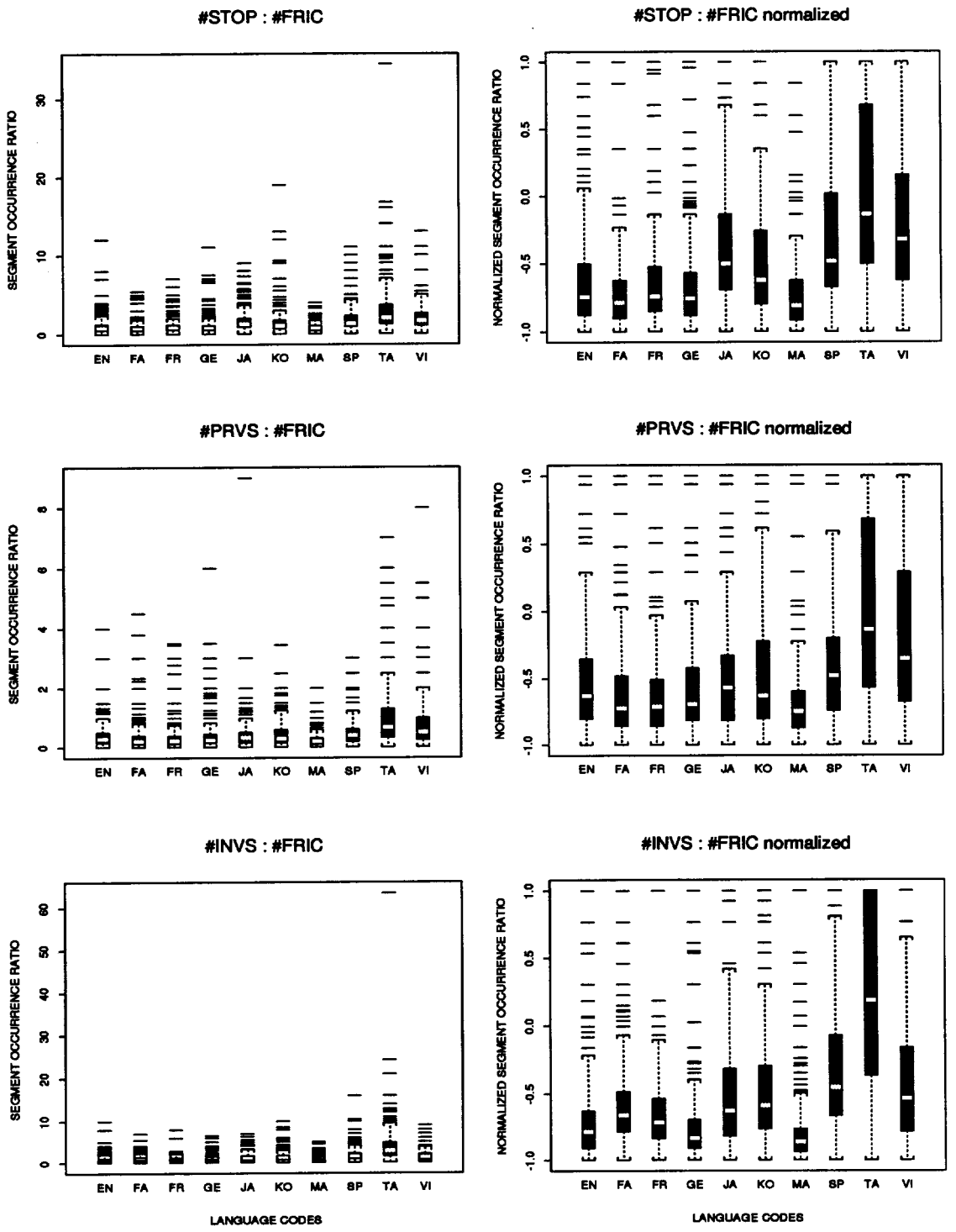

Figure C.84: Boxplots of Segment Occurrence Ratios (continued)

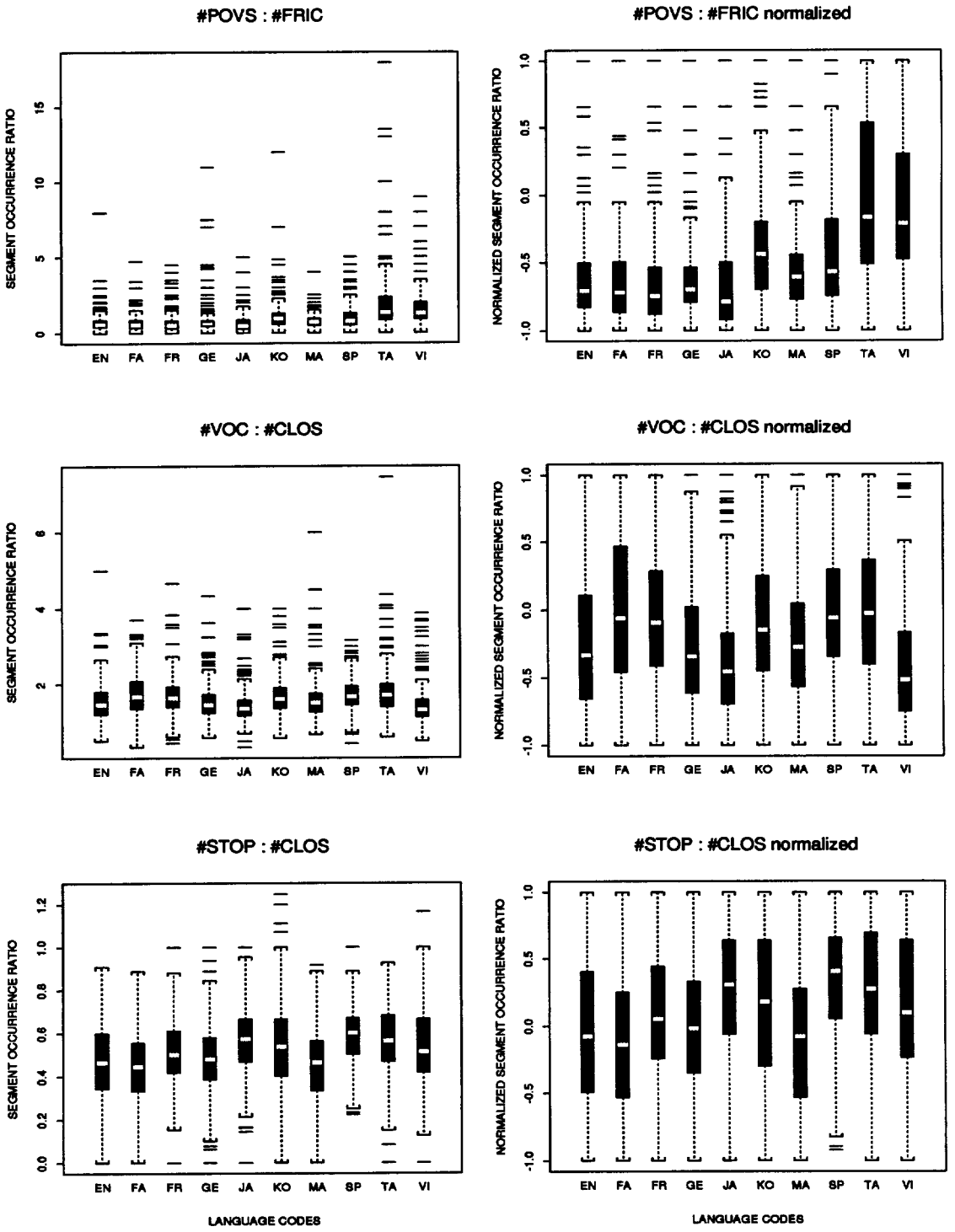

Figure C.85: Boxplots of Segment Occurrence Ratios (continued)

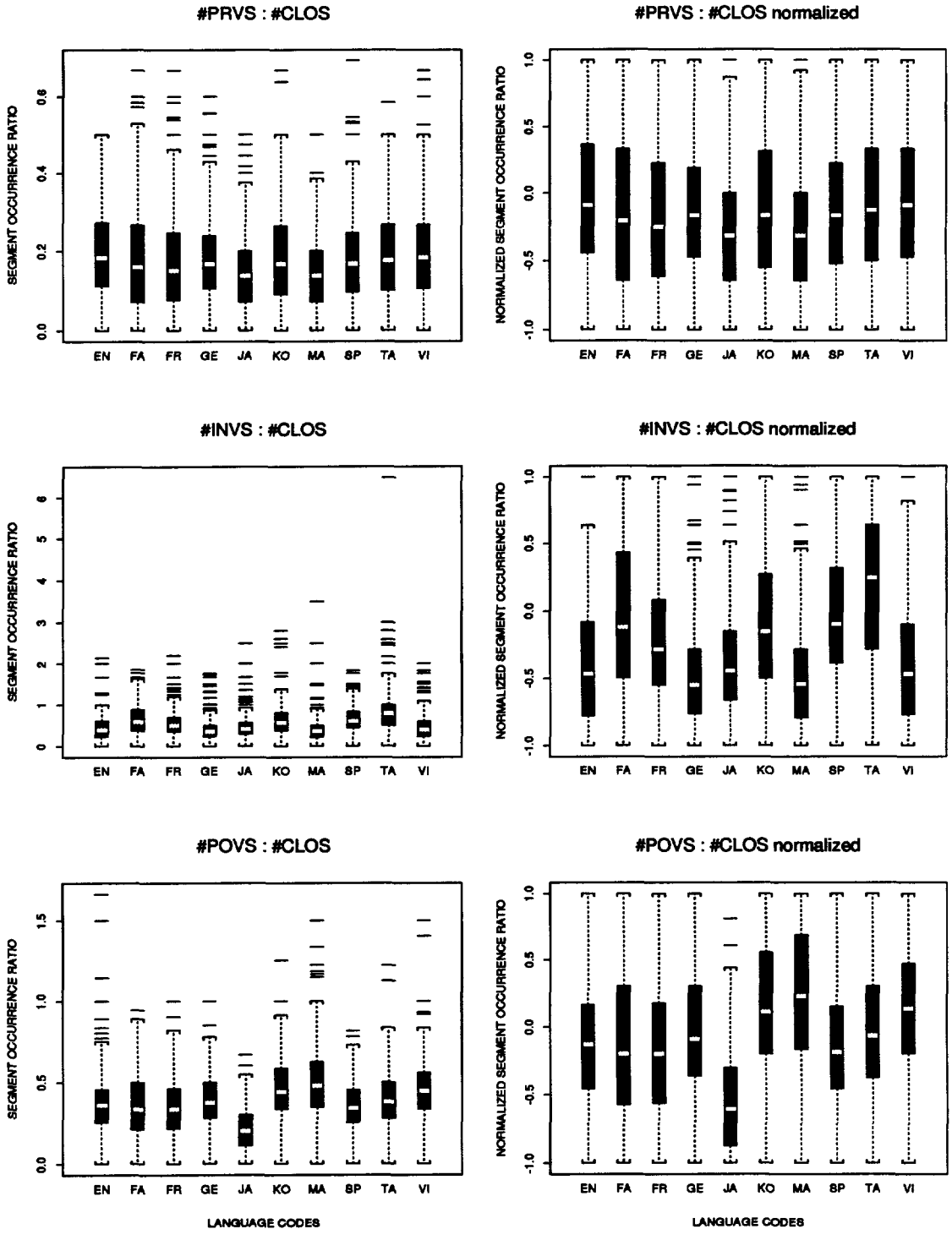

Figure C.86: Boxplots of Segment Occurrence Ratios (continued)

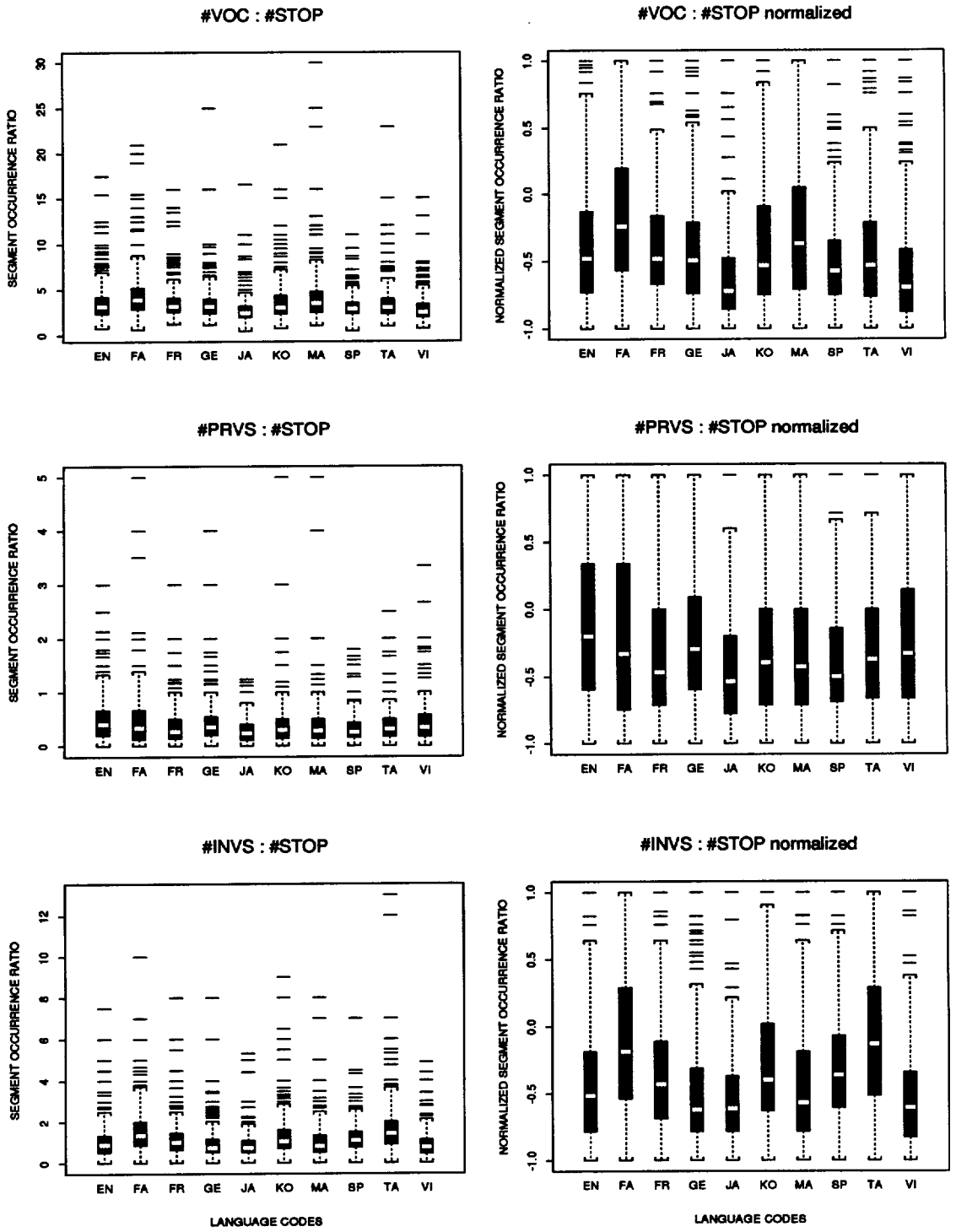

Figure C.87: Boxplots of Segment Occurrence Ratios (continued)

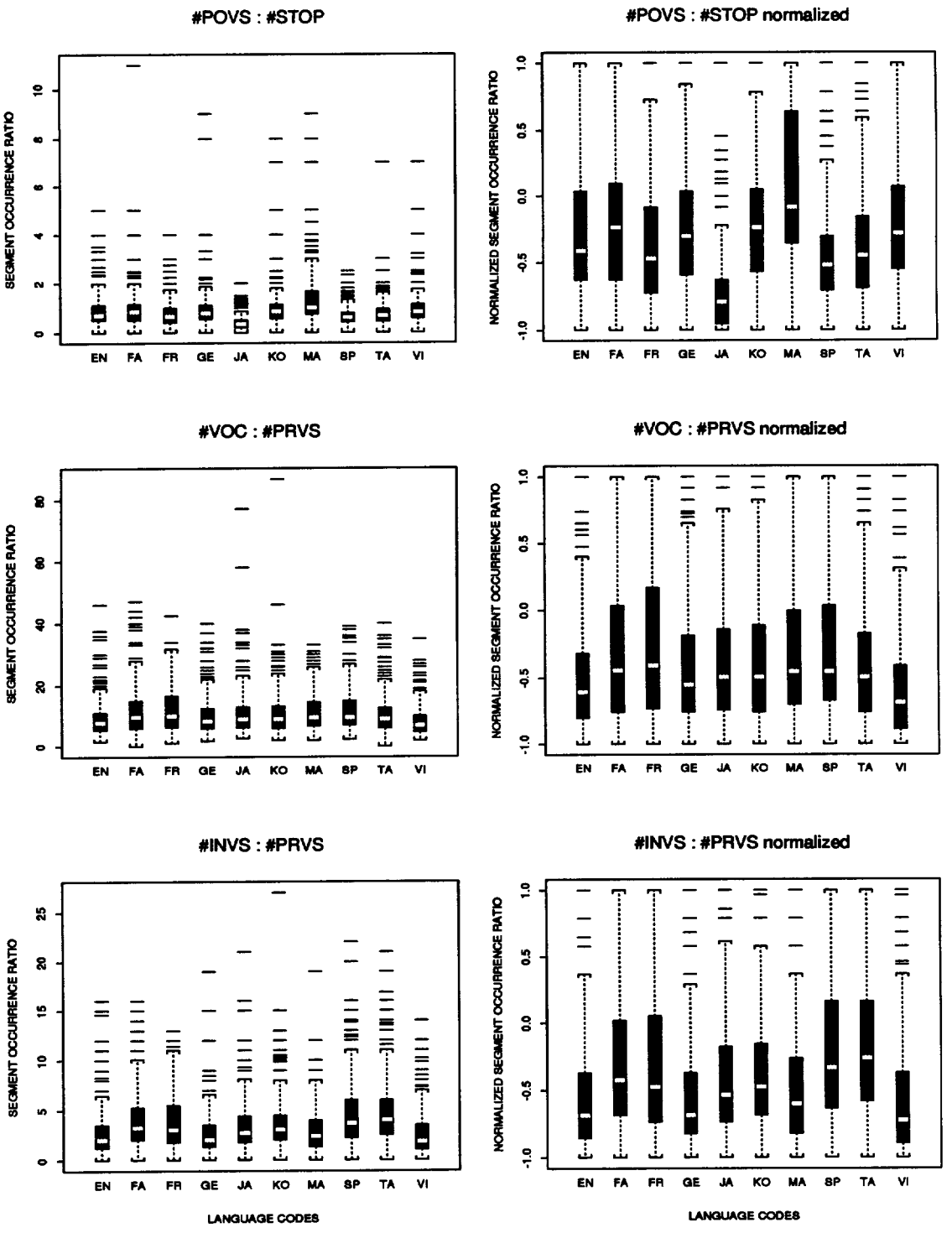

Figure C.88: Boxplots of Segment Occurrence Ratios (continued)

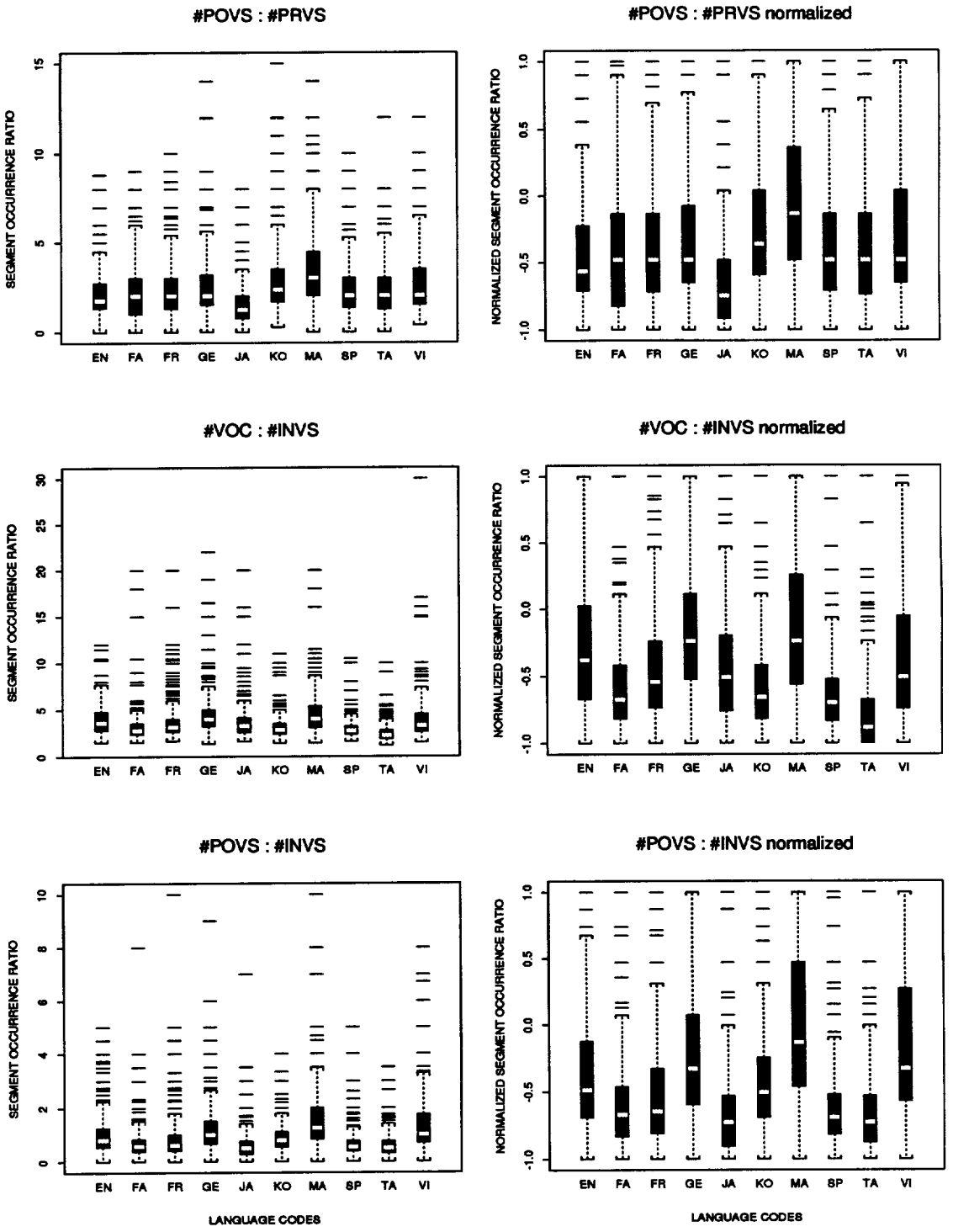

Figure C.89: Boxplots of Segment Occurrence Ratios (continued)

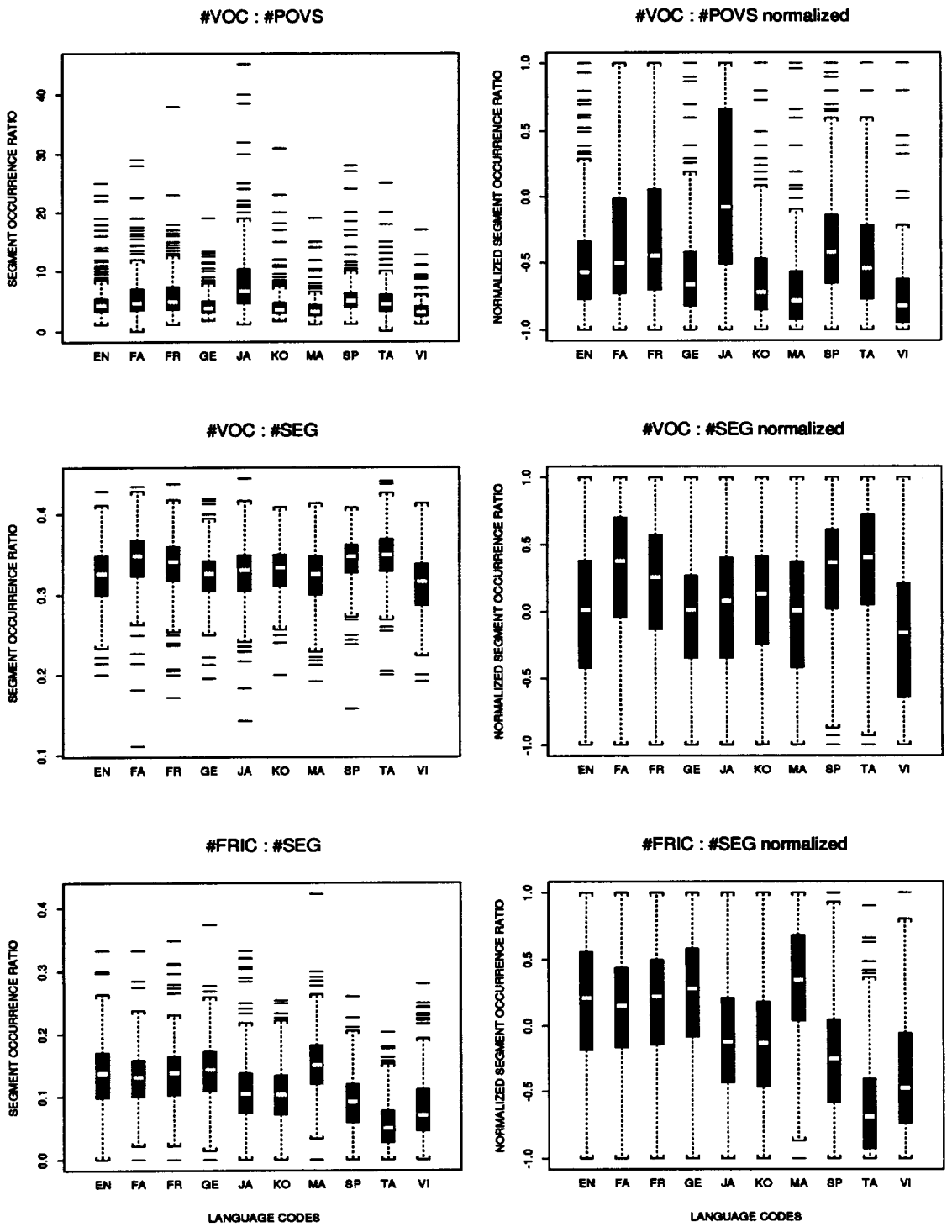

Figure C.90: Boxplots of Segment Occurrence Ratios (continued)

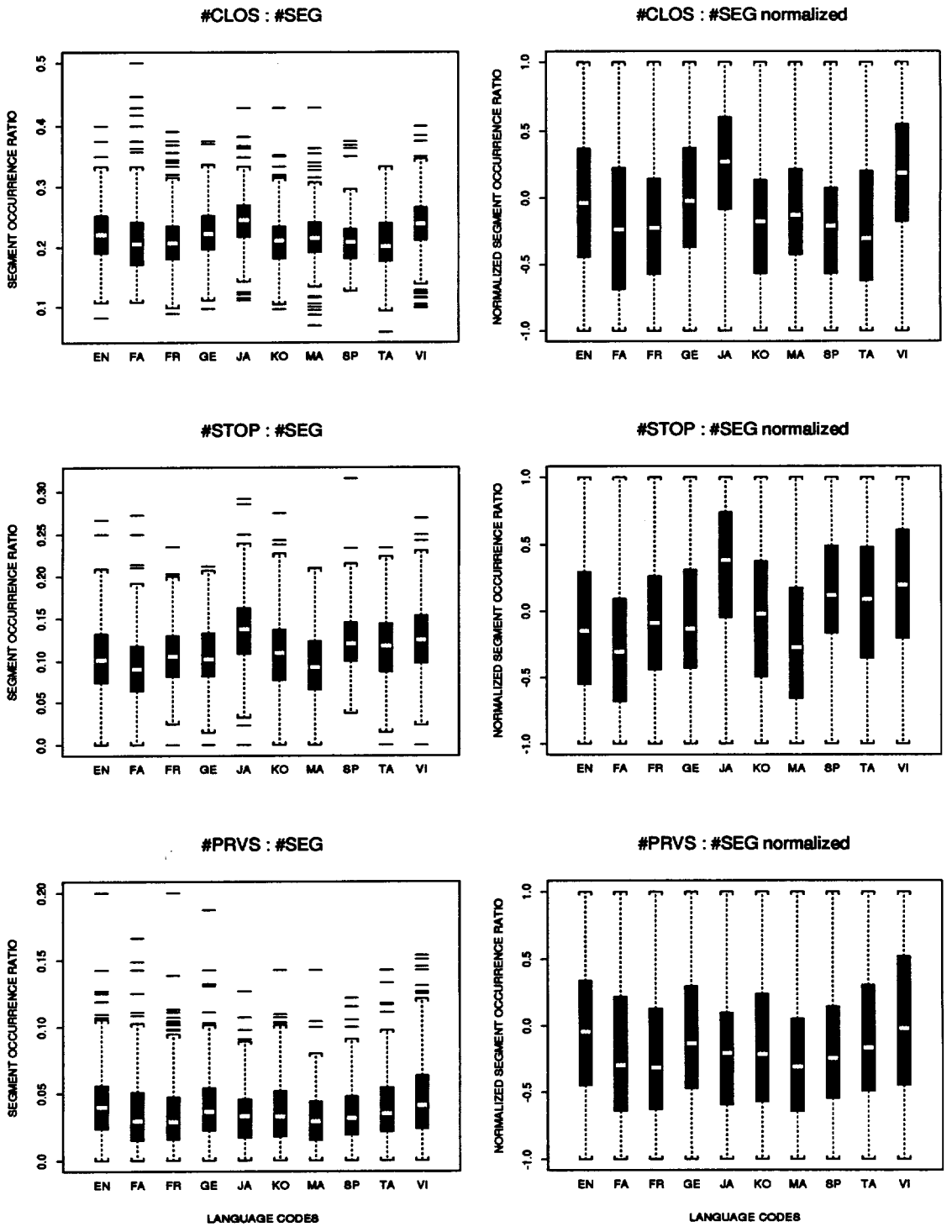

Figure C.91: Boxplots of Segment Occurrence Ratios (continued)

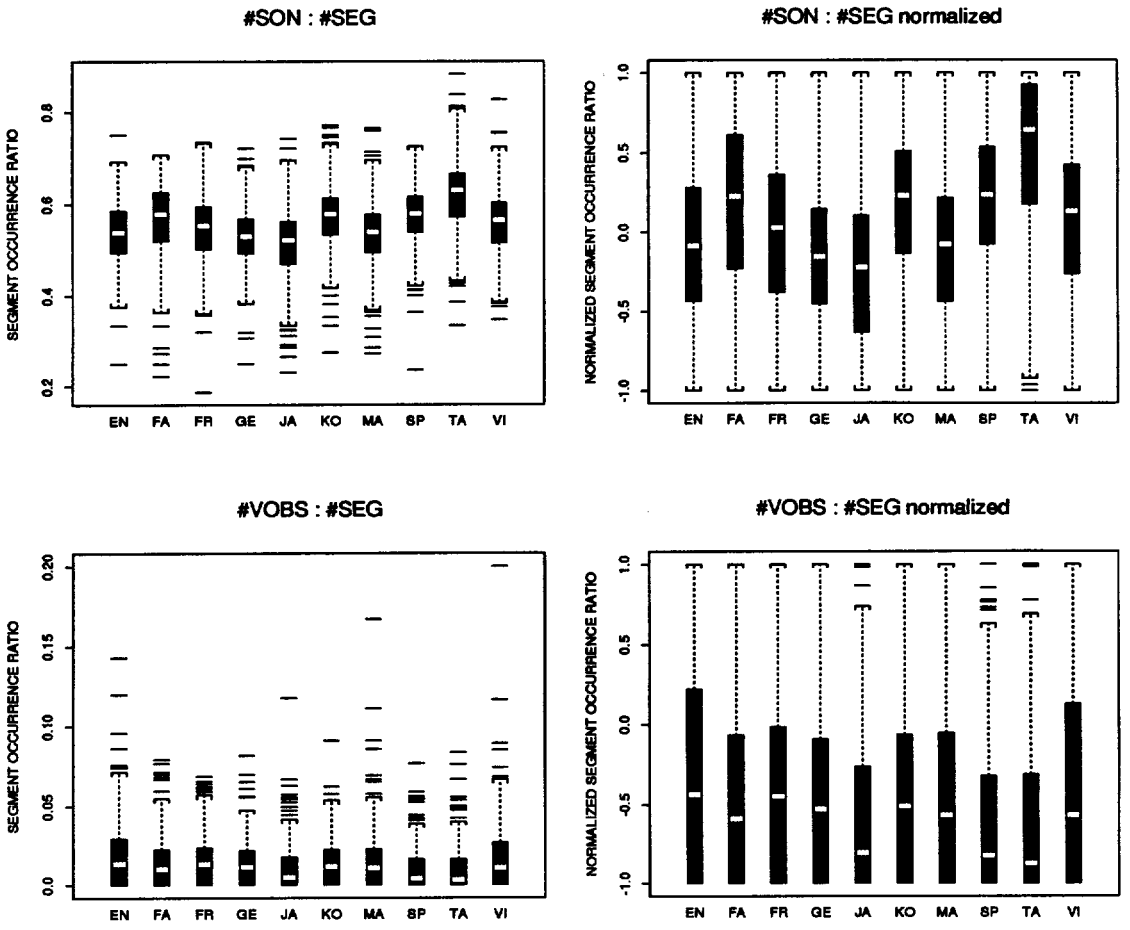

Figure C.92: Boxplots of Segment Occurrence Ratios (continued)

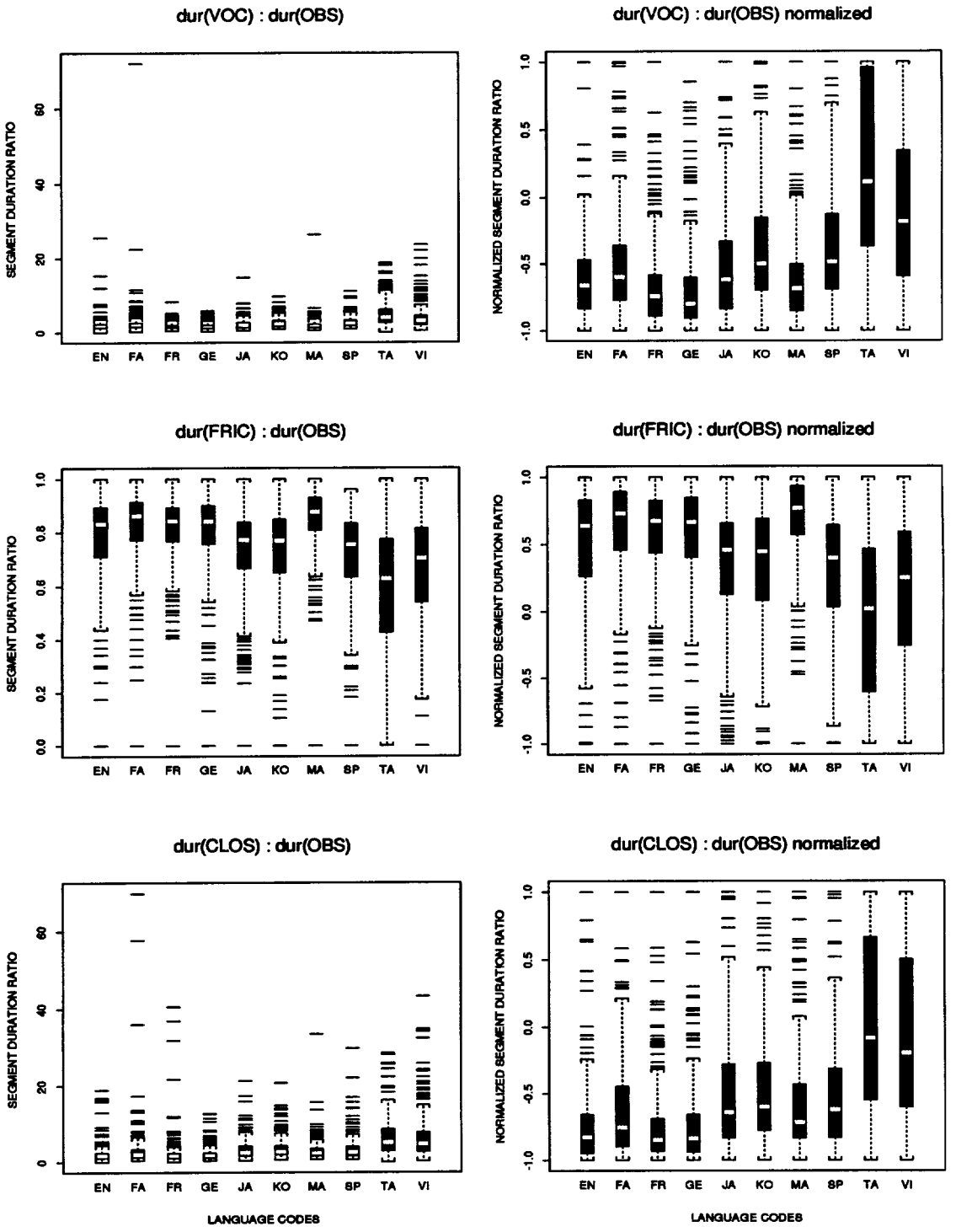

Figure C.93: Boxplots of Segment Duration Ratios

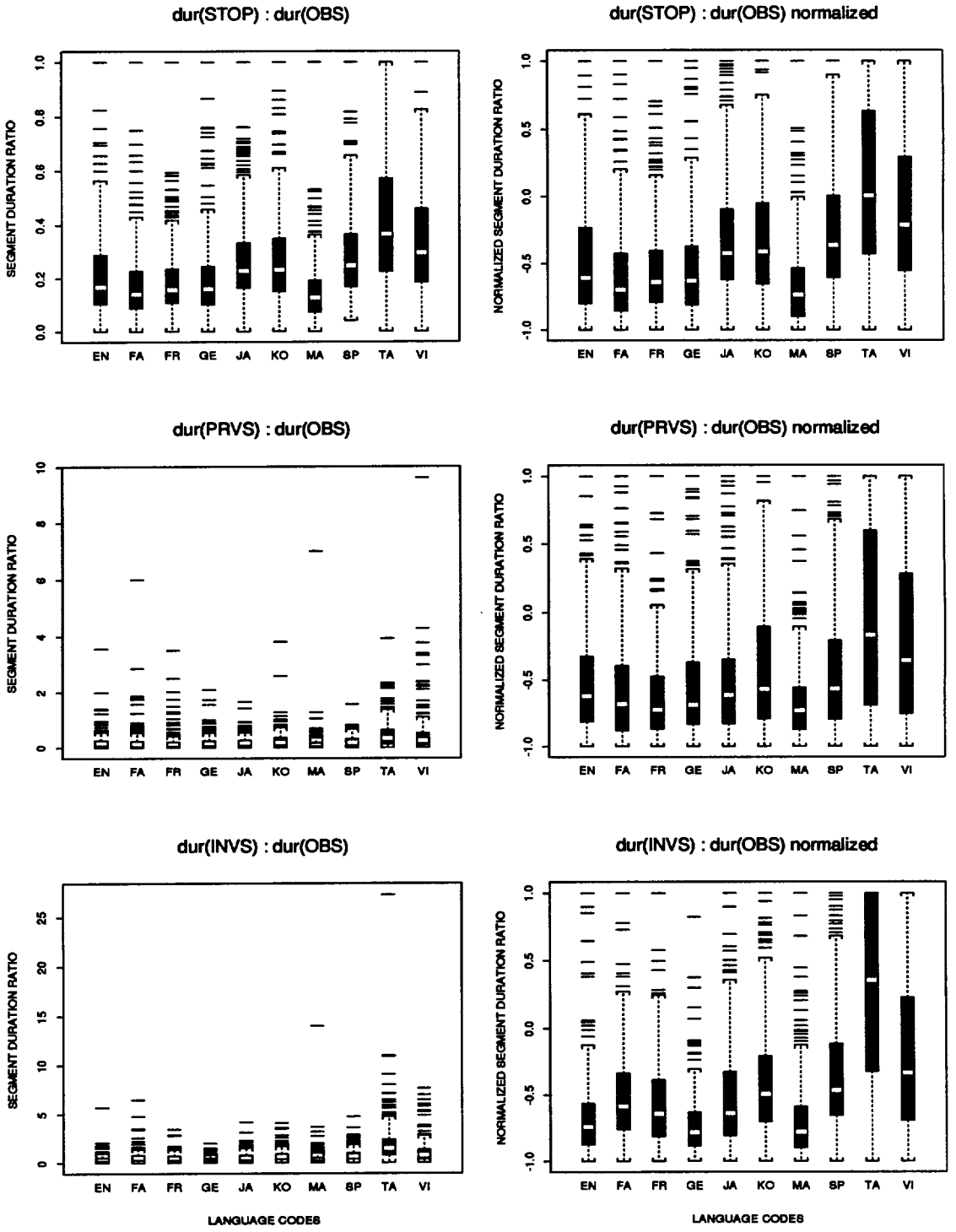

Figure C.94: Boxplots of Segment Duration Ratios (continued)

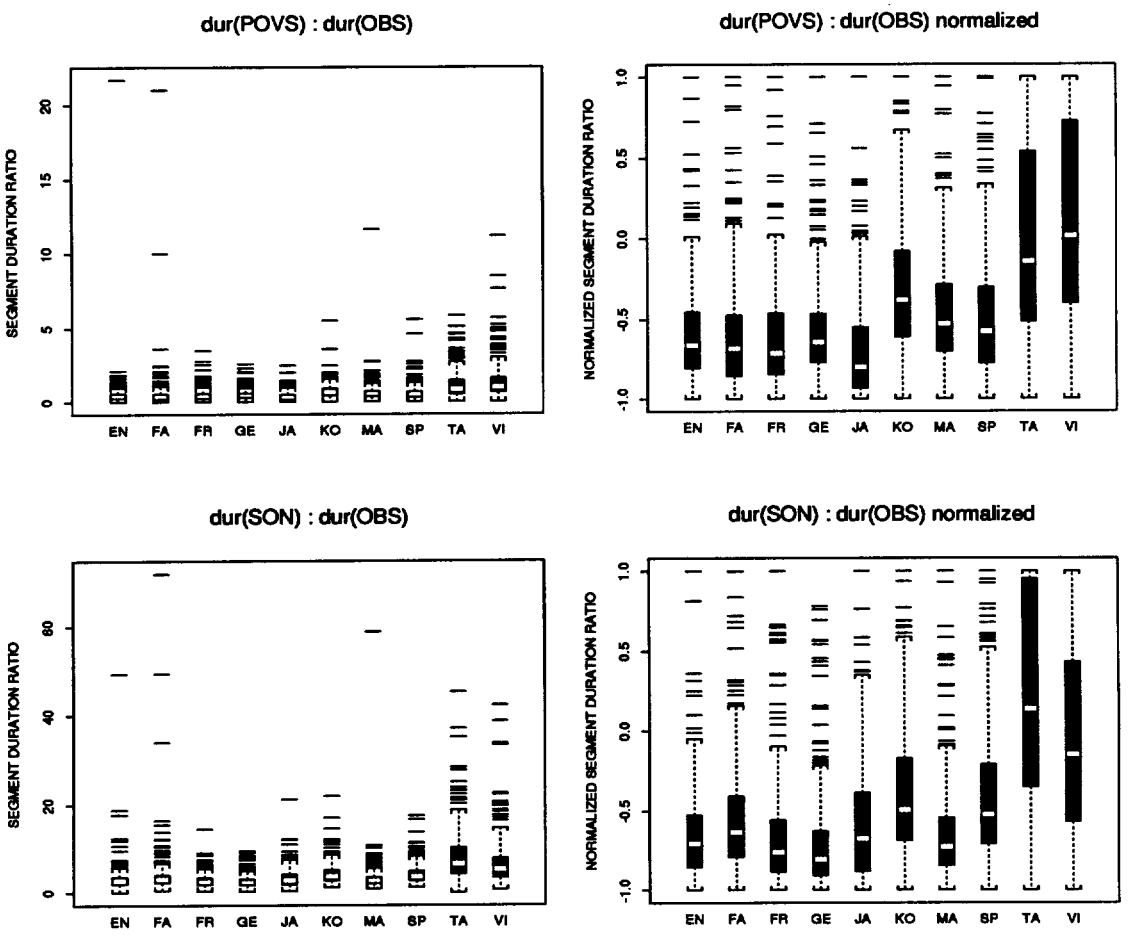

Figure C.95: Boxplots of Segment Duration Ratios (continued)

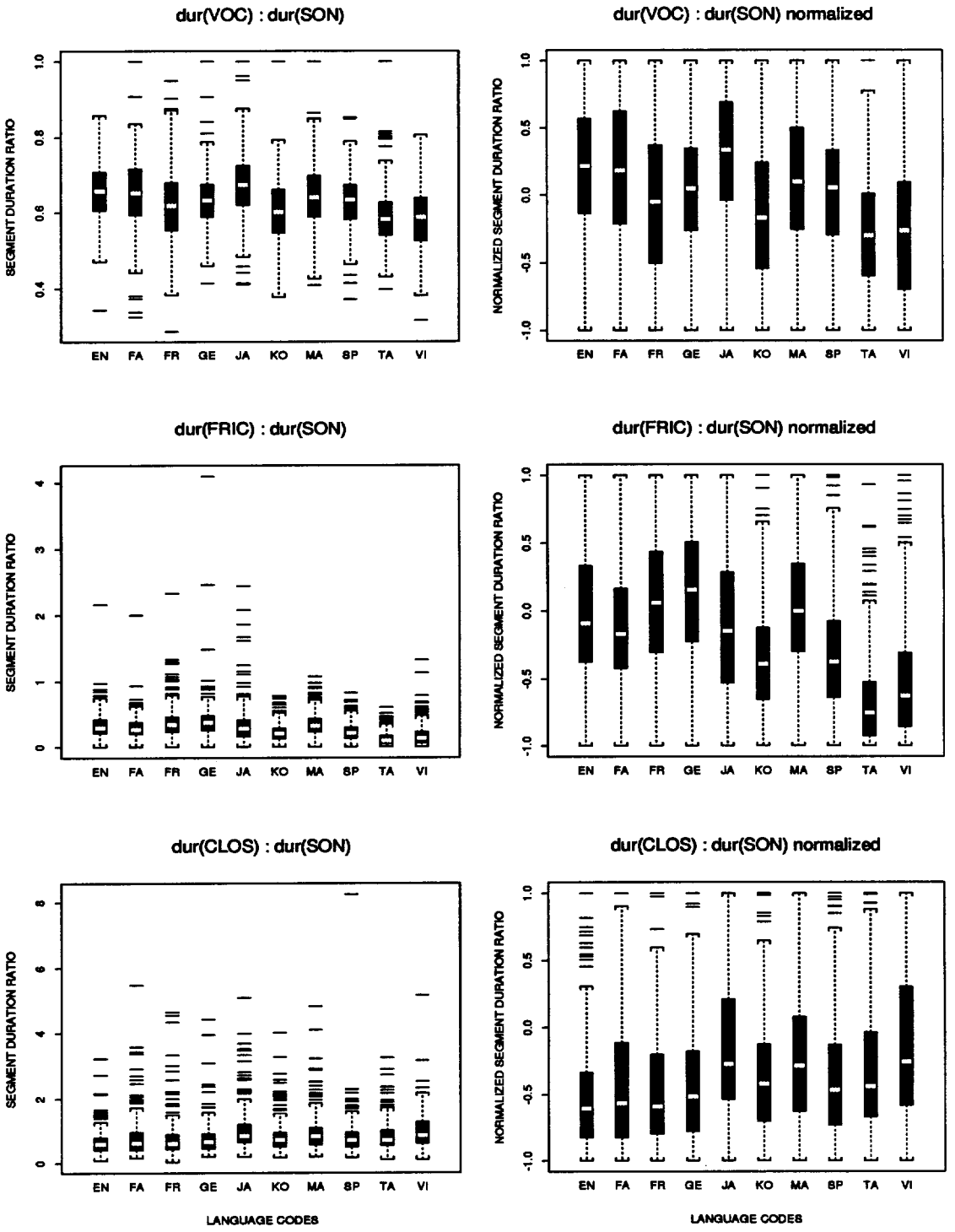

Figure C.96: Boxplots of Segment Duration Ratios (continued)

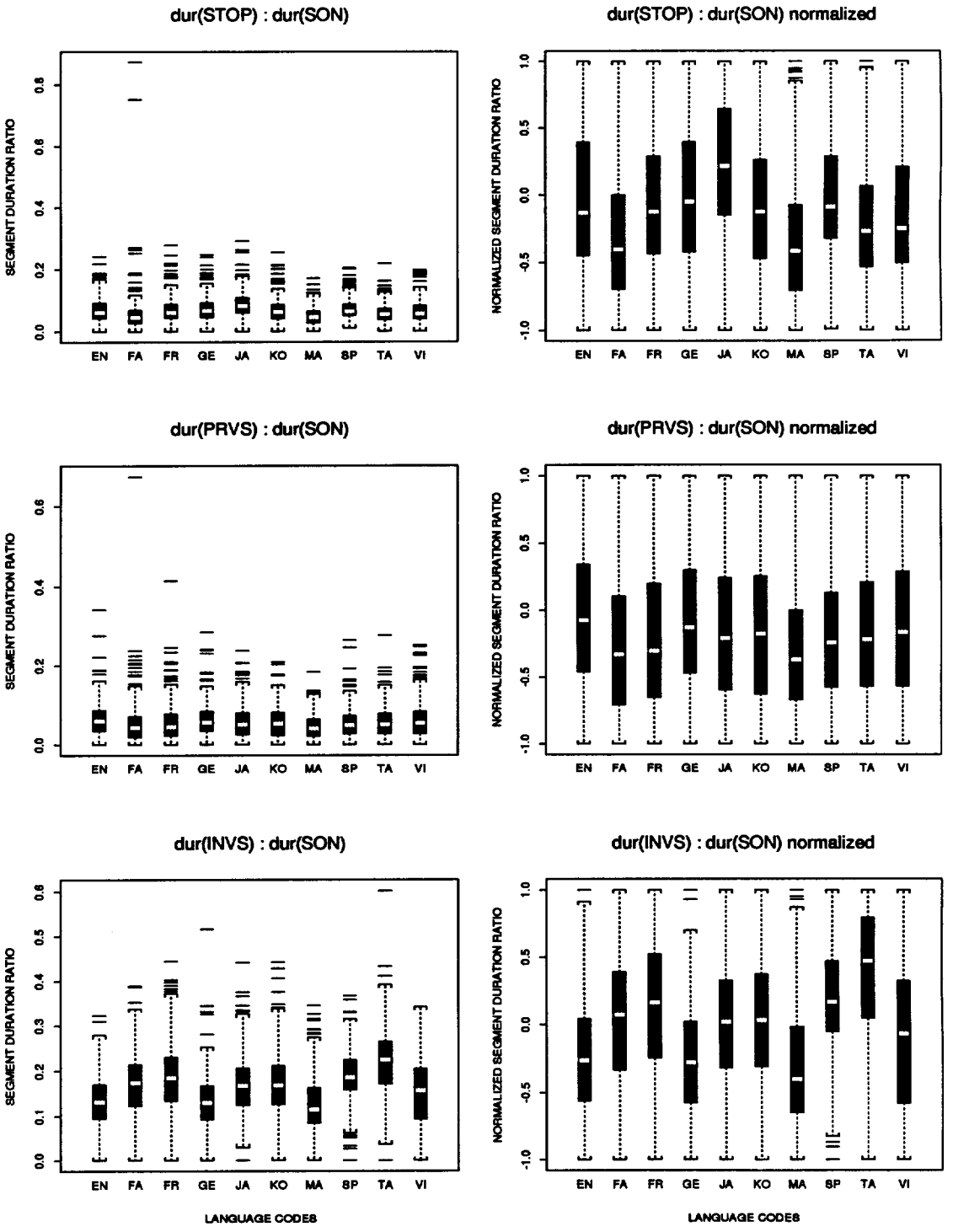

Figure C.97: Boxplots of Segment Duration Ratios (continued)

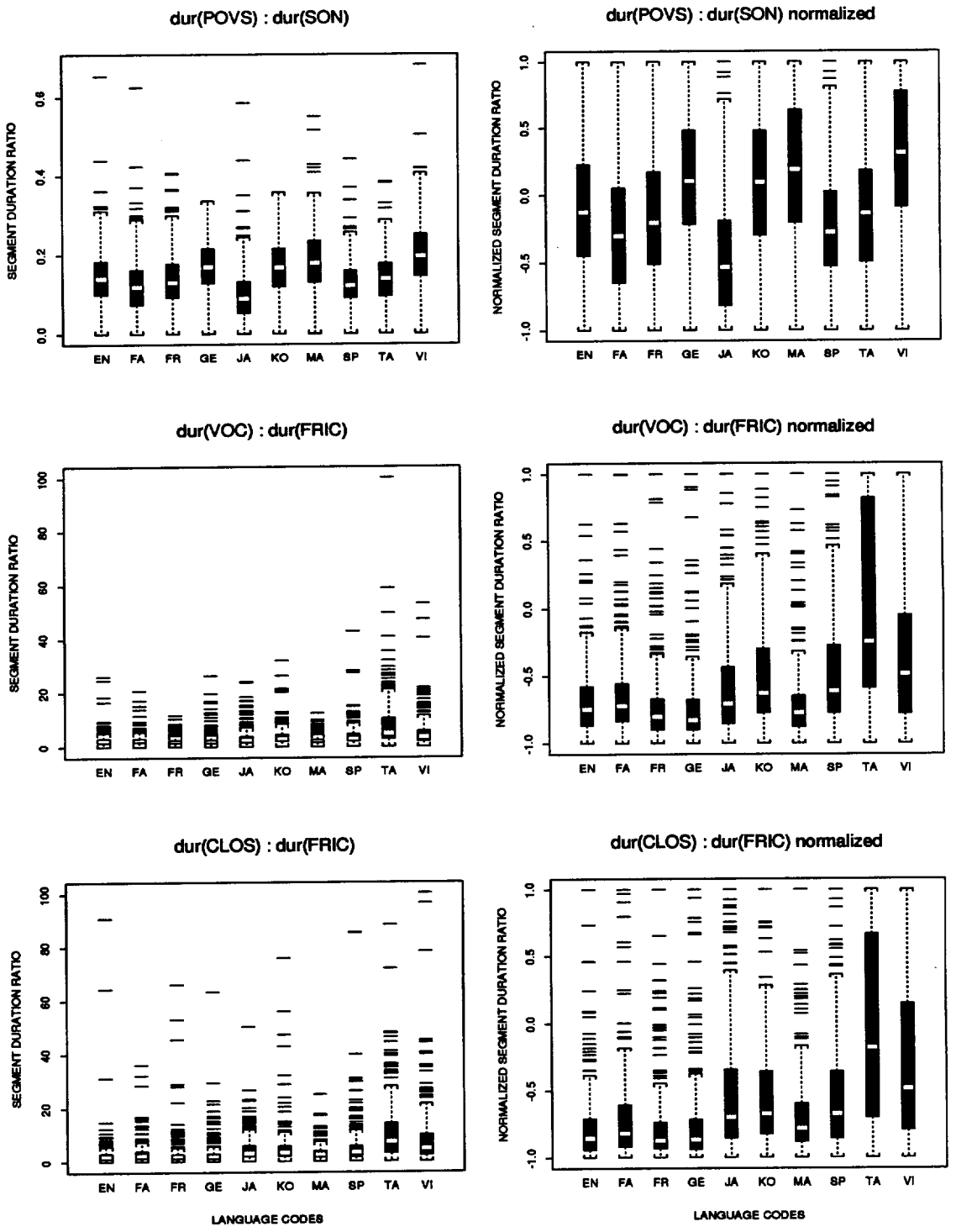

Figure C.98: Boxplots of Segment Duration Ratios (continued)

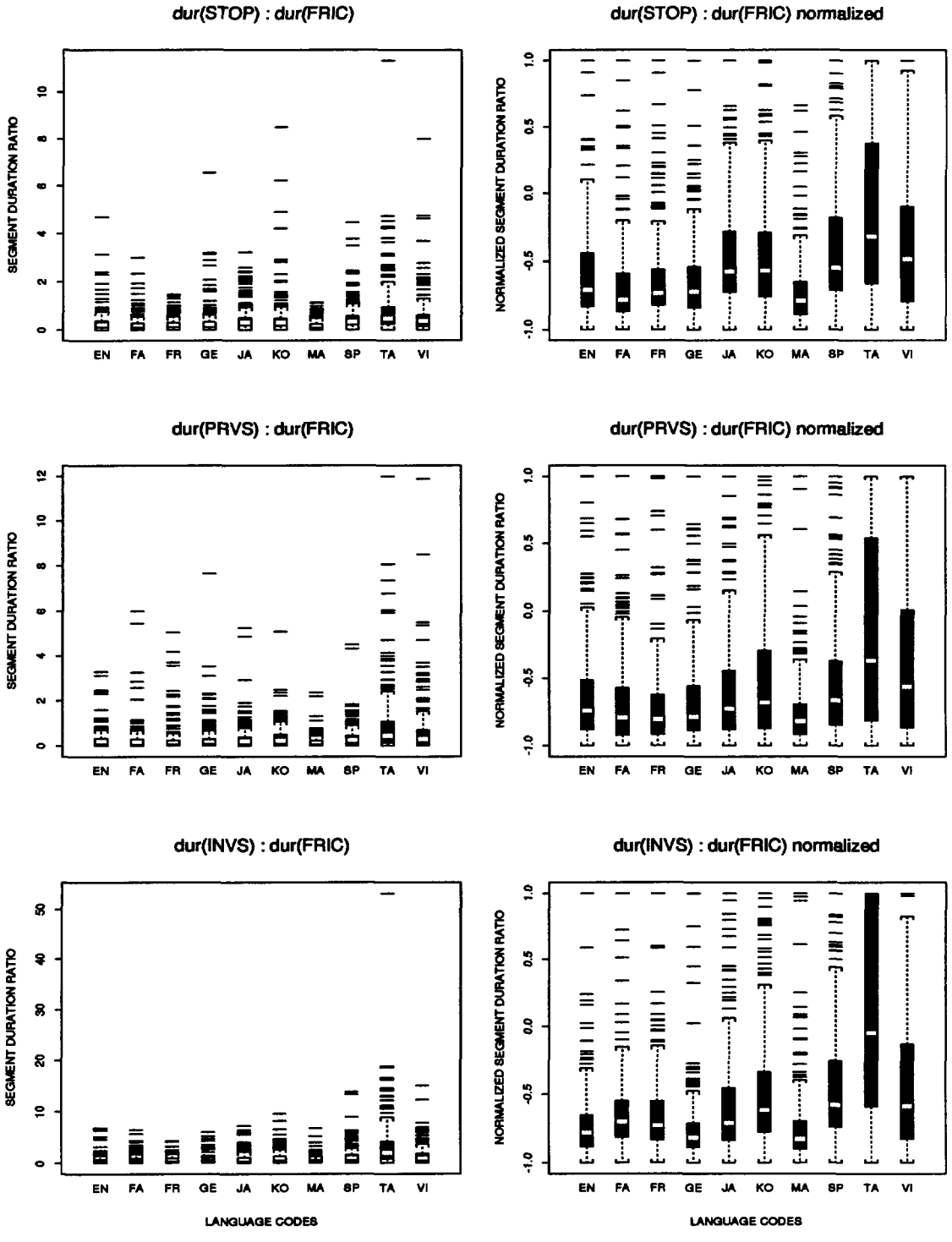

Figure C.99: Boxplots of Segment Duration Ratios (continued)

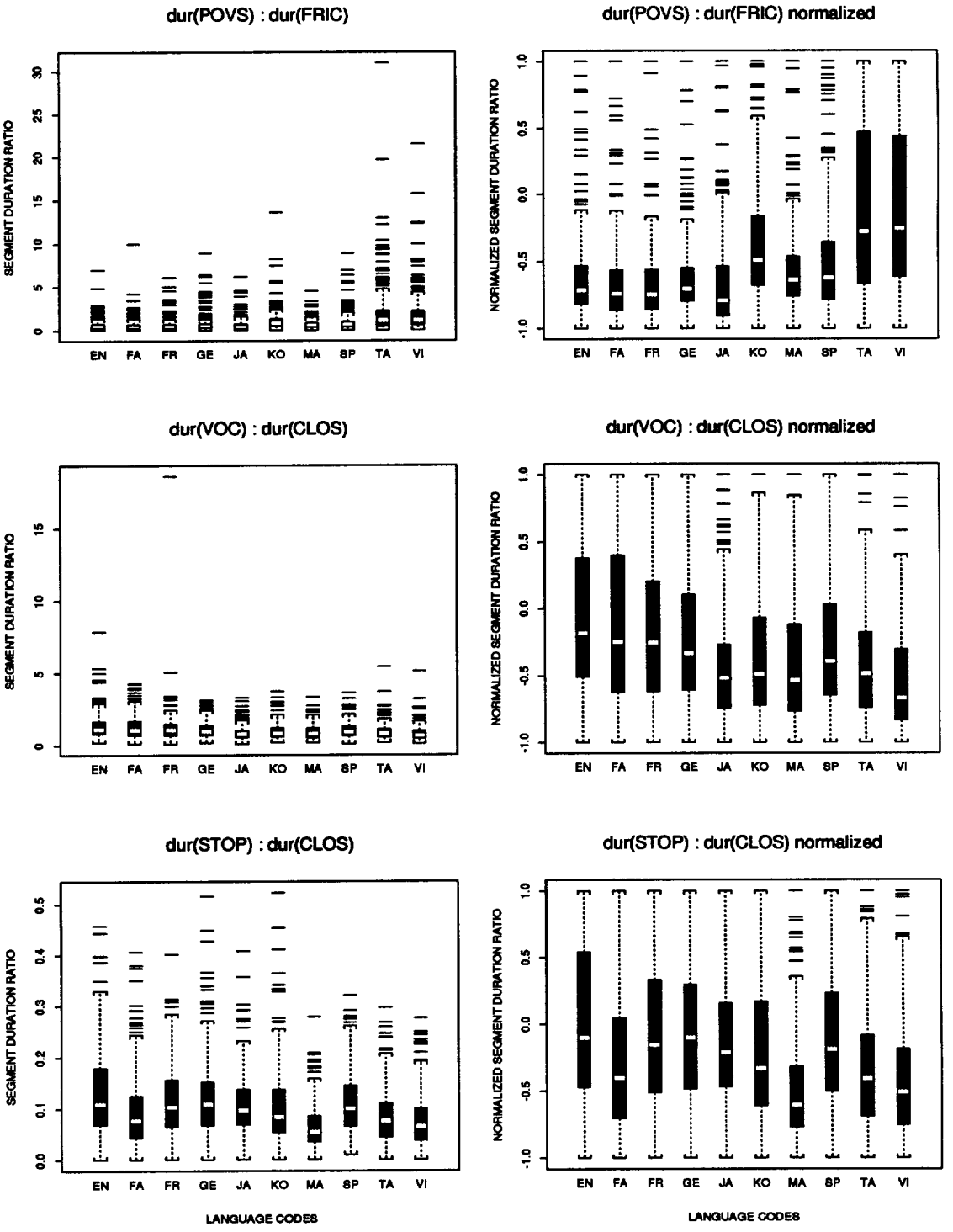

Figure C.100: Boxplots of Segment Duration Ratios (continued)

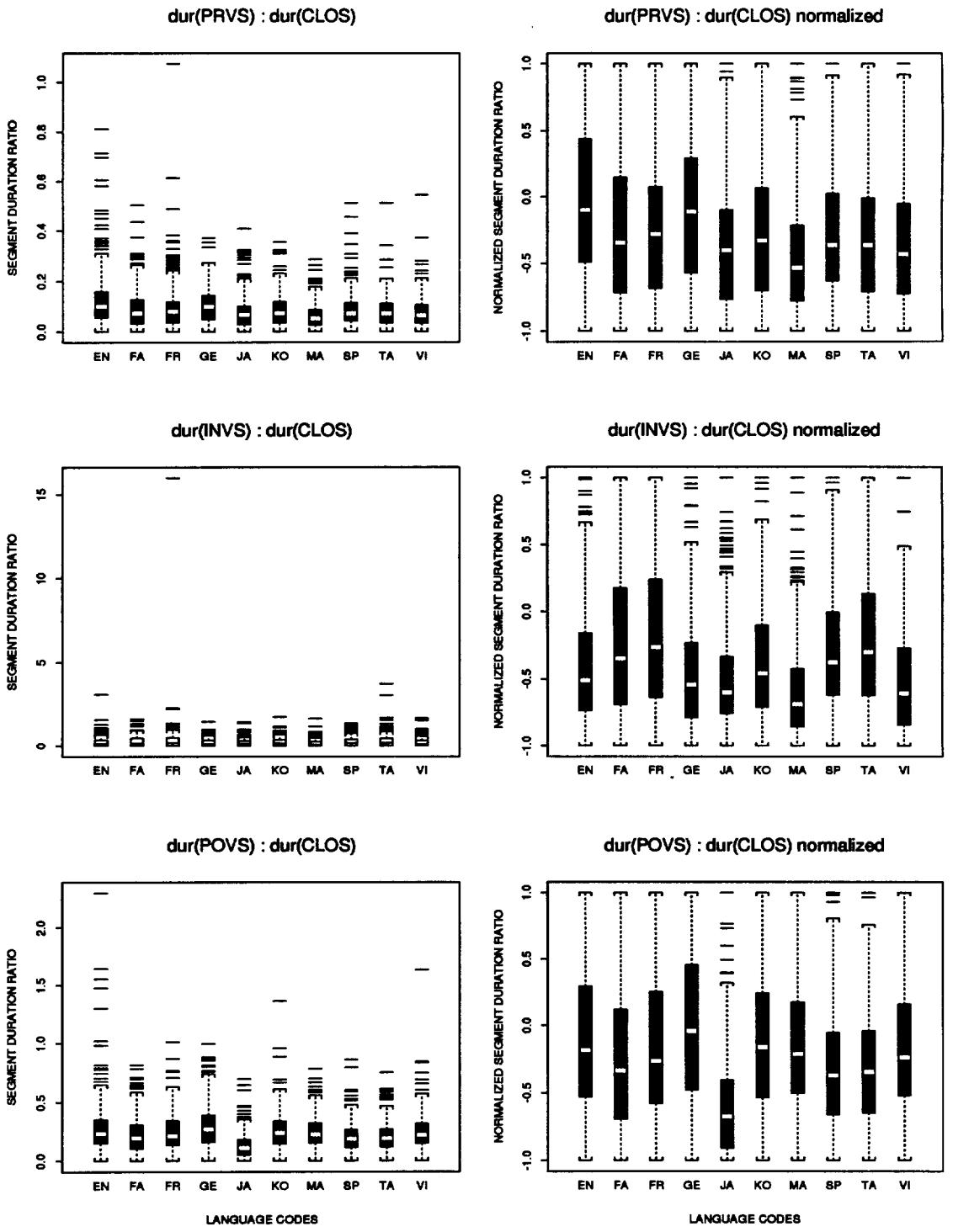

Figure C.101: Boxplots of Segment Duration Ratios (continued)

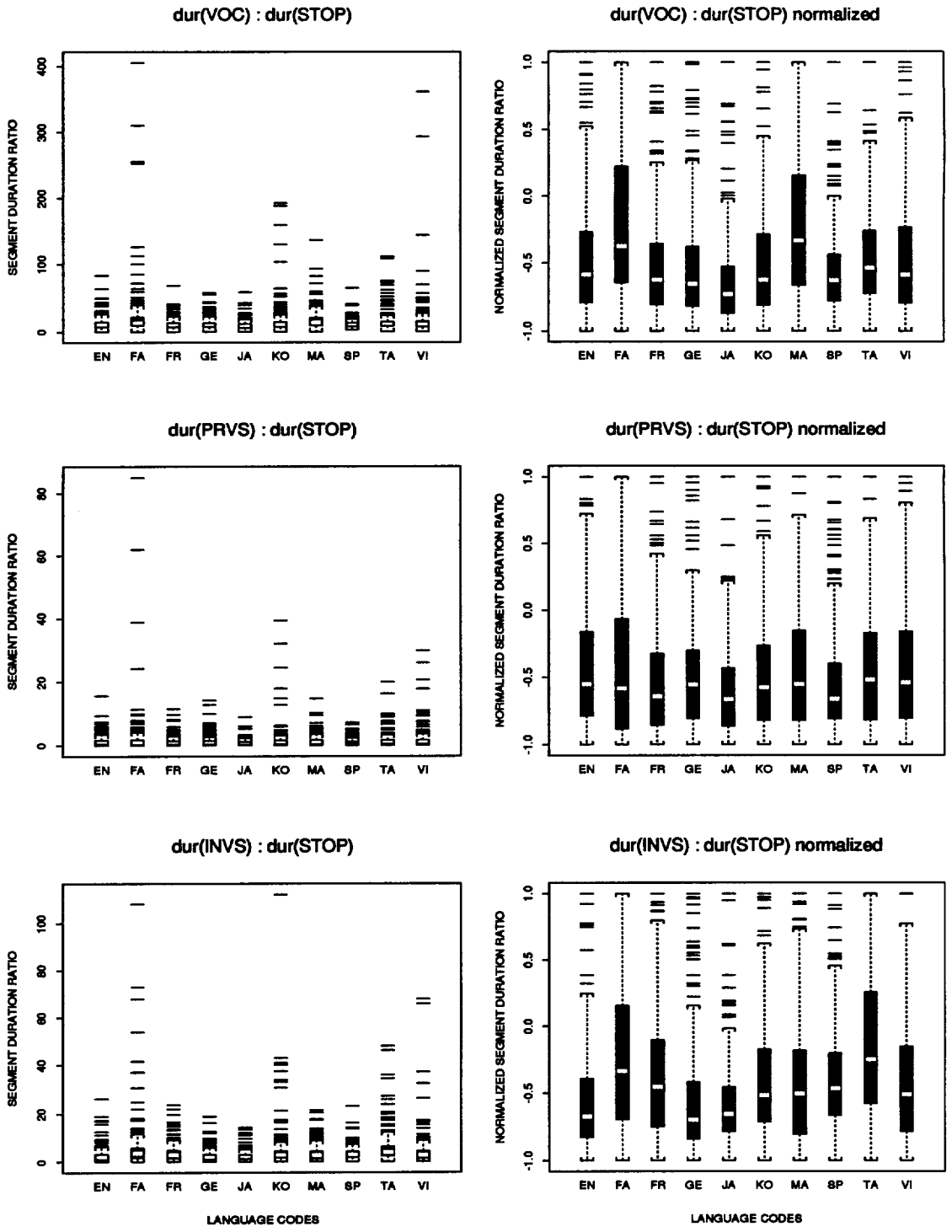

Figure C.102: Boxplots of Segment Duration Ratios (continued)

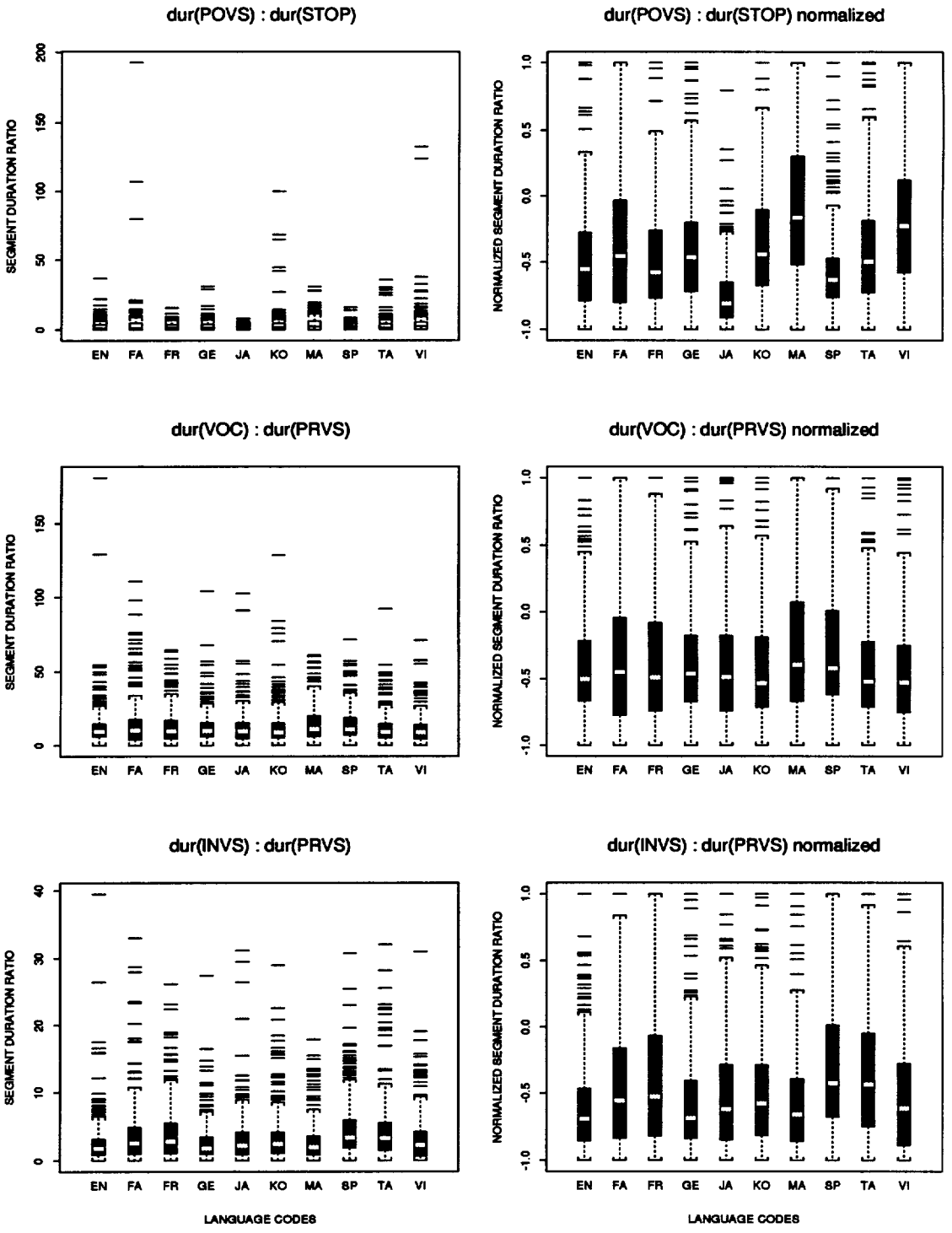

Figure C.103: Boxplots of Segment Duration Ratios (continued)

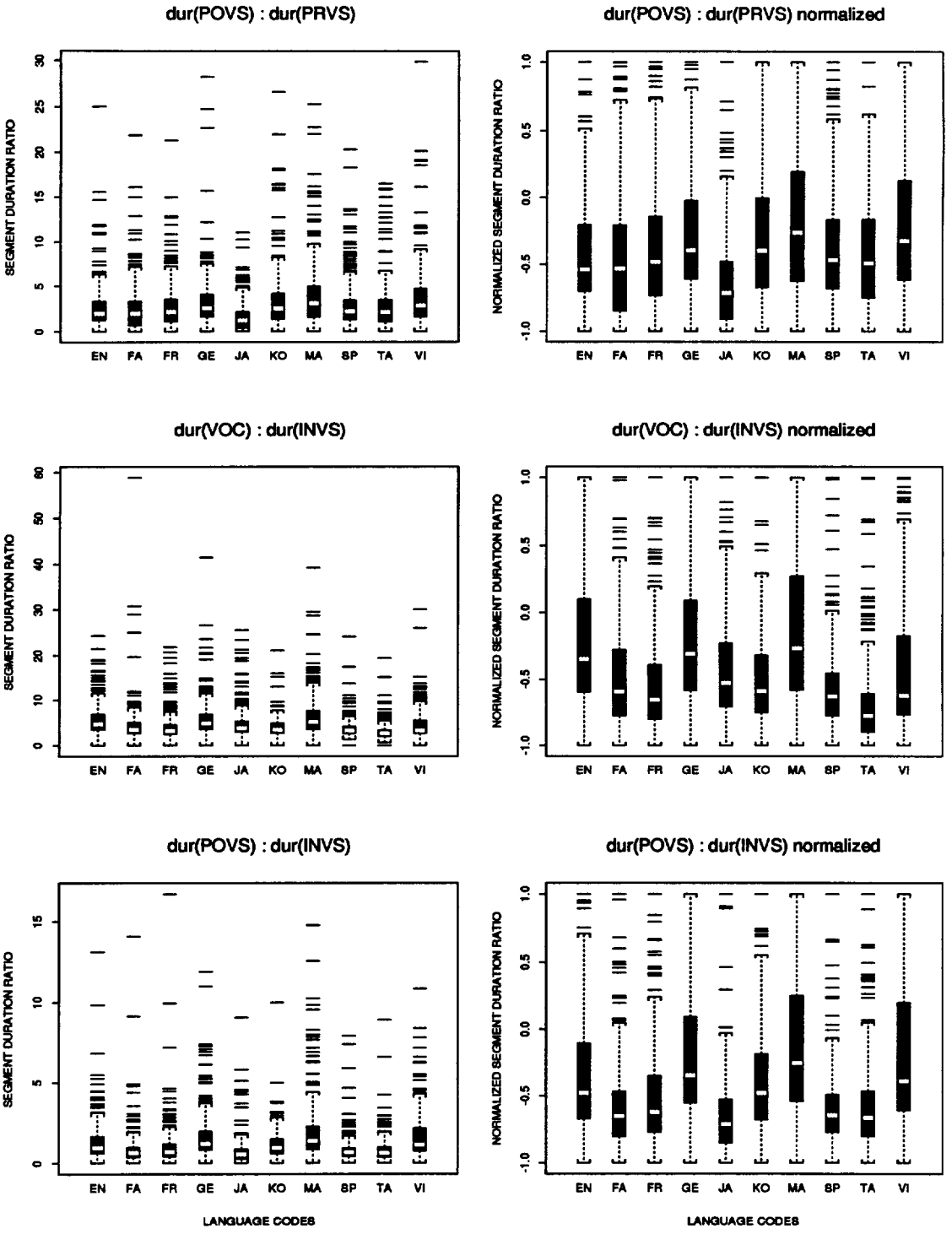

Figure C.104: Boxplots of Segment Duration Ratios (continued)

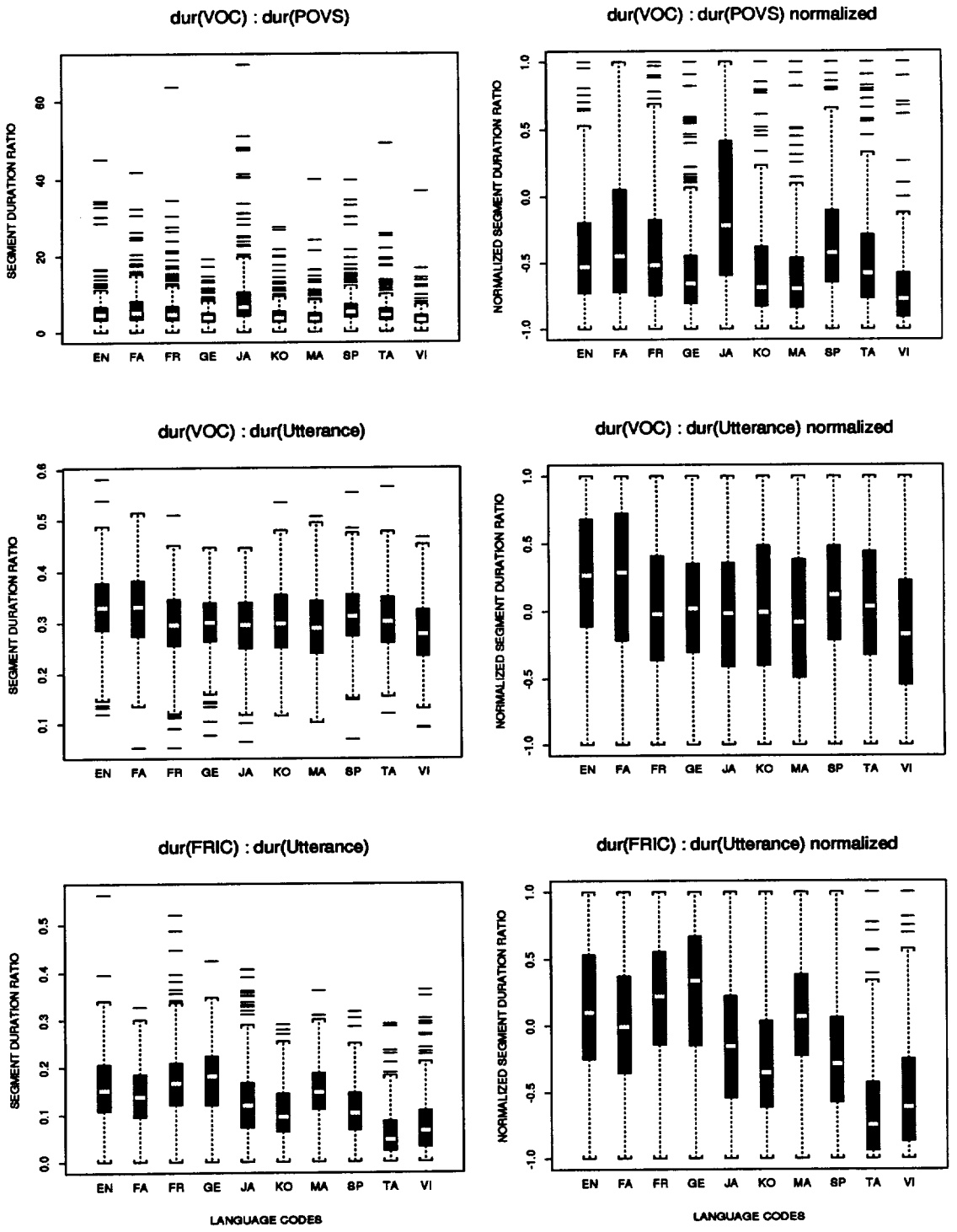

Figure C.105: Boxplots of Segment Duration Ratios (continued)

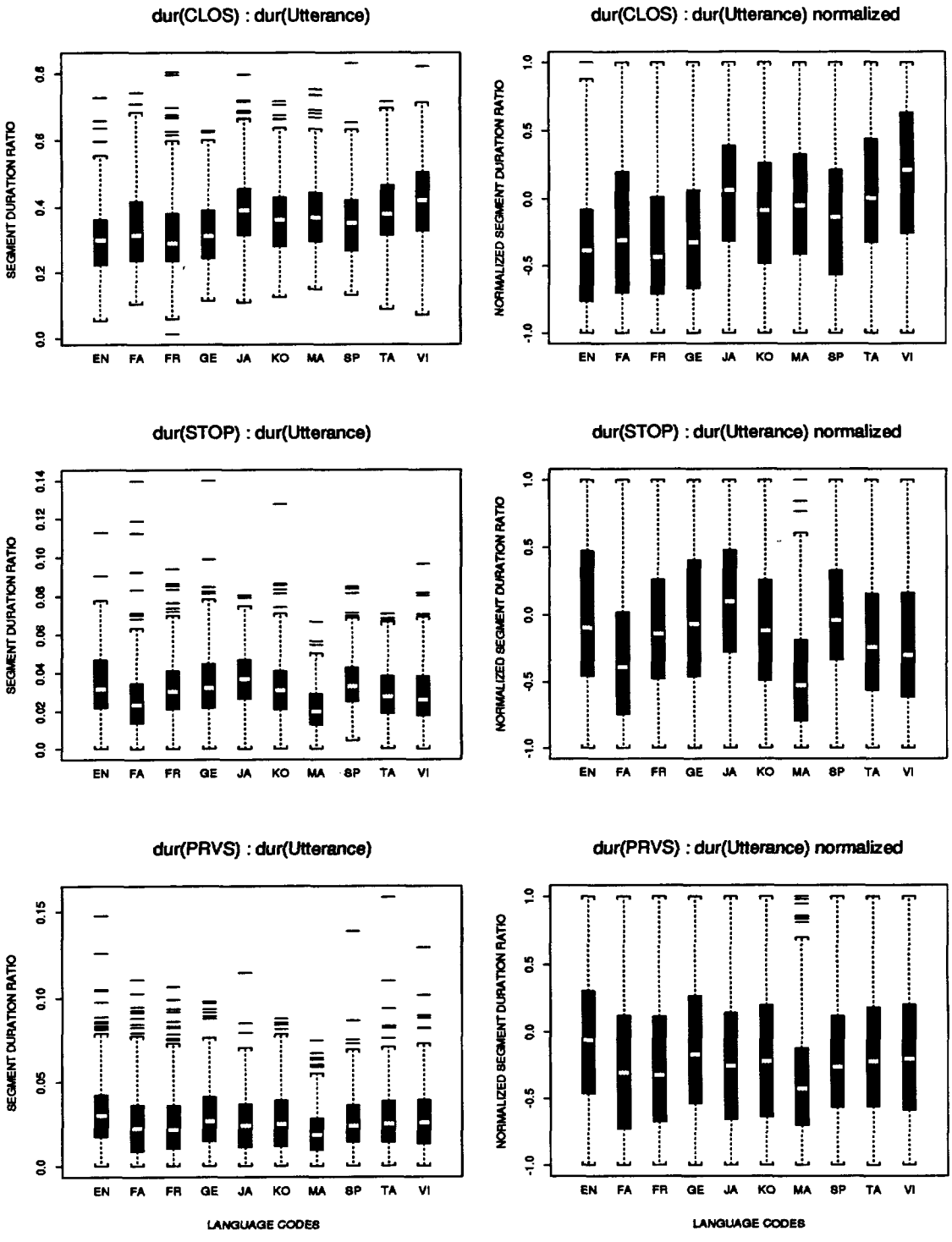

Figure C.106: Boxplots of Segment Duration Ratios (continued)

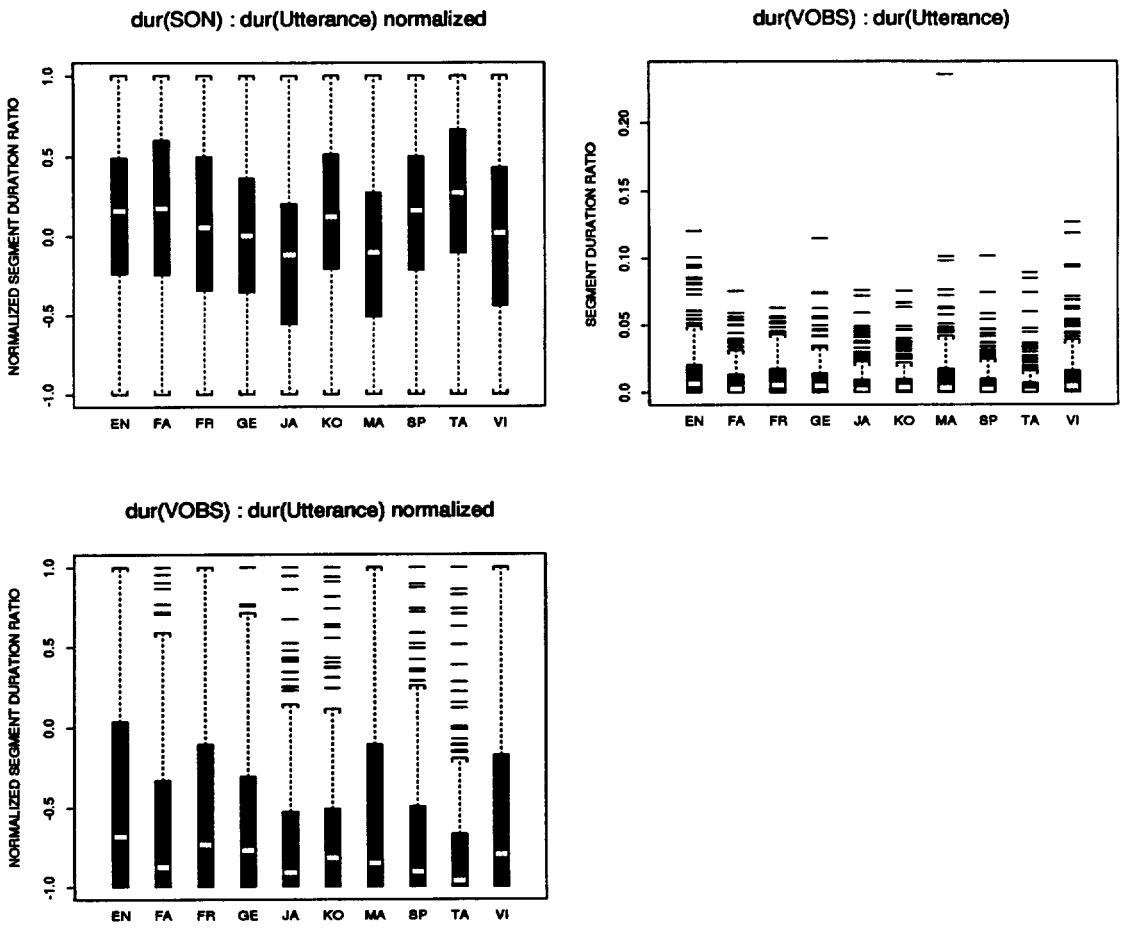

Figure C.107: Boxplots of Segment Duration Ratios (continued)

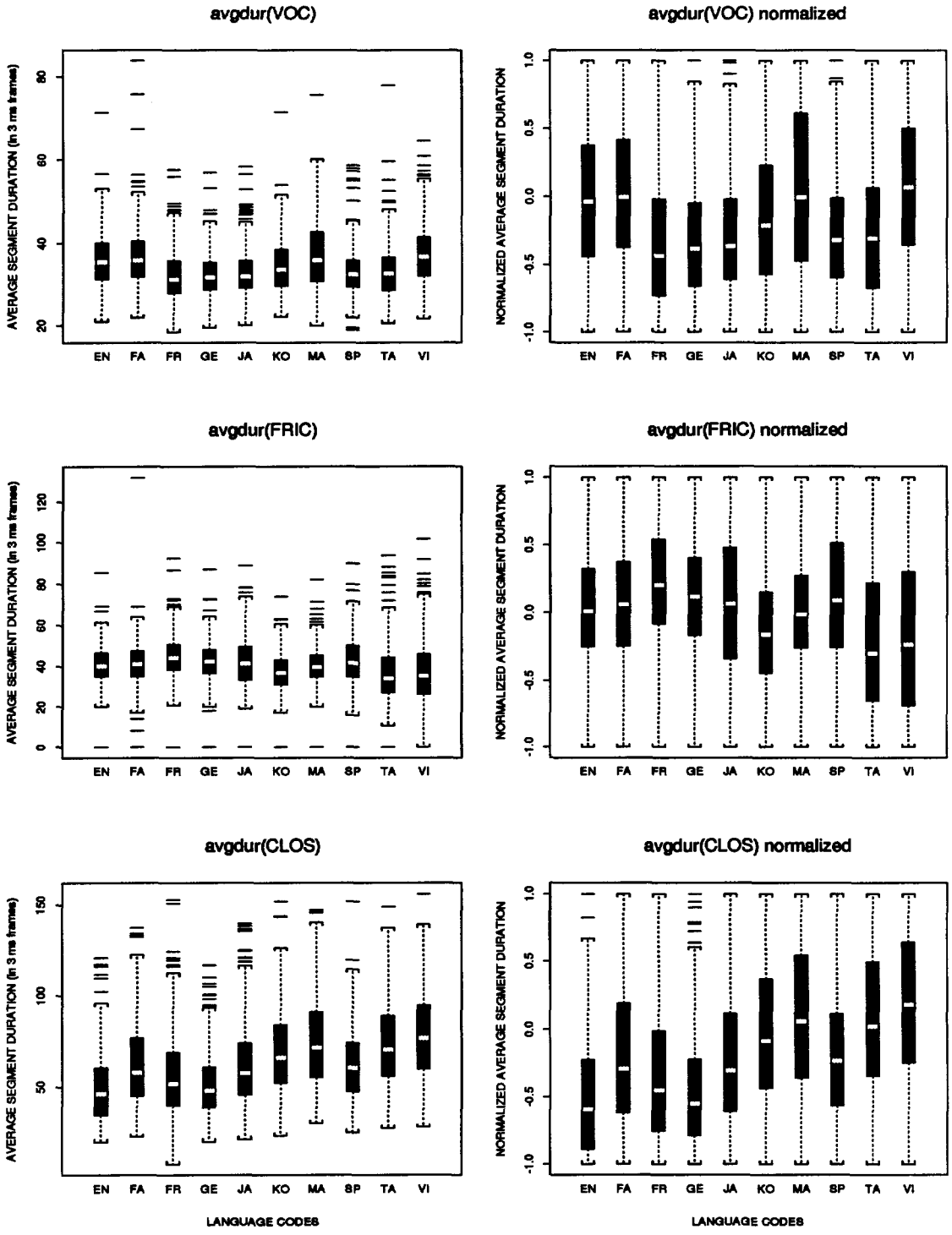

Figure C.108: Average Duration of VOC, FRIC and CLOS

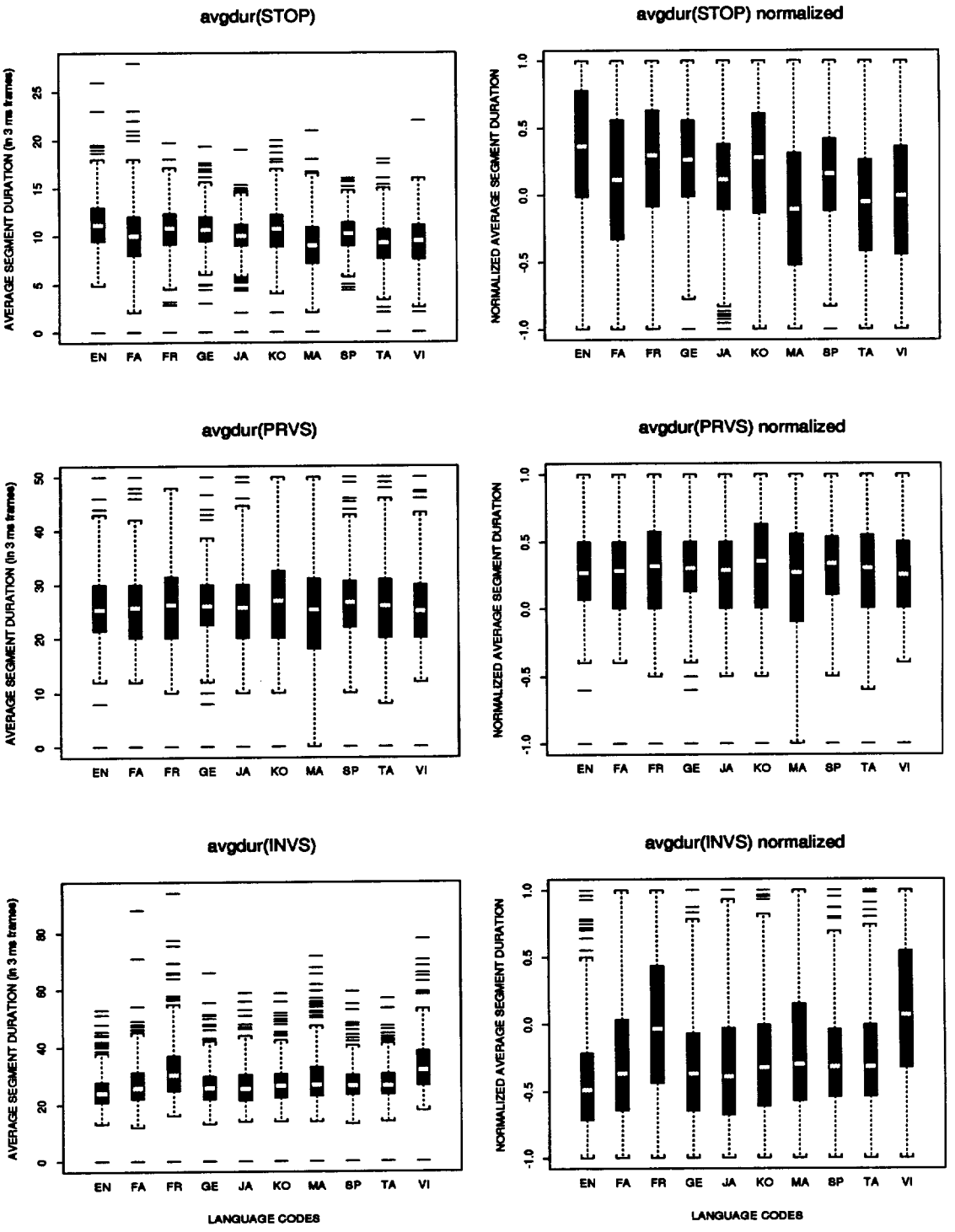

Figure C.109: Average Duration of STOP, PRVS and INVS

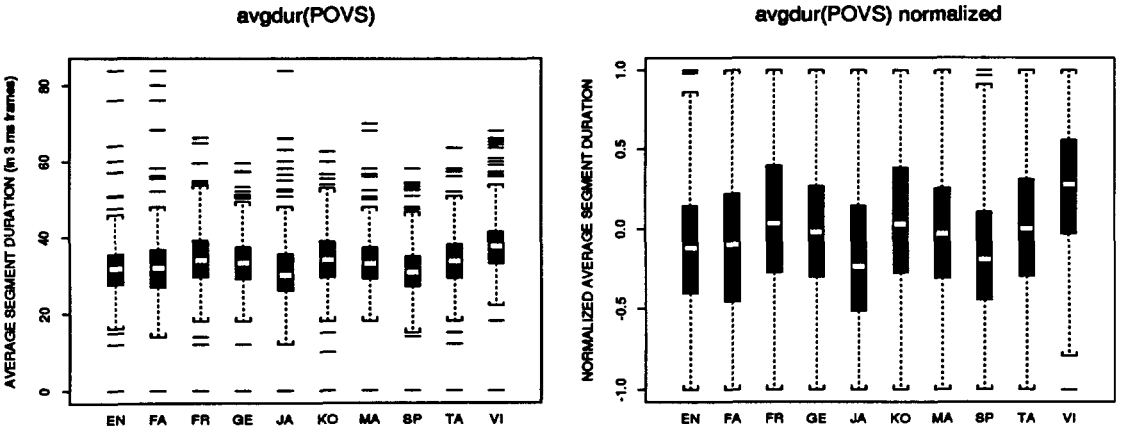

Figure C.110: Average Duration of POVS

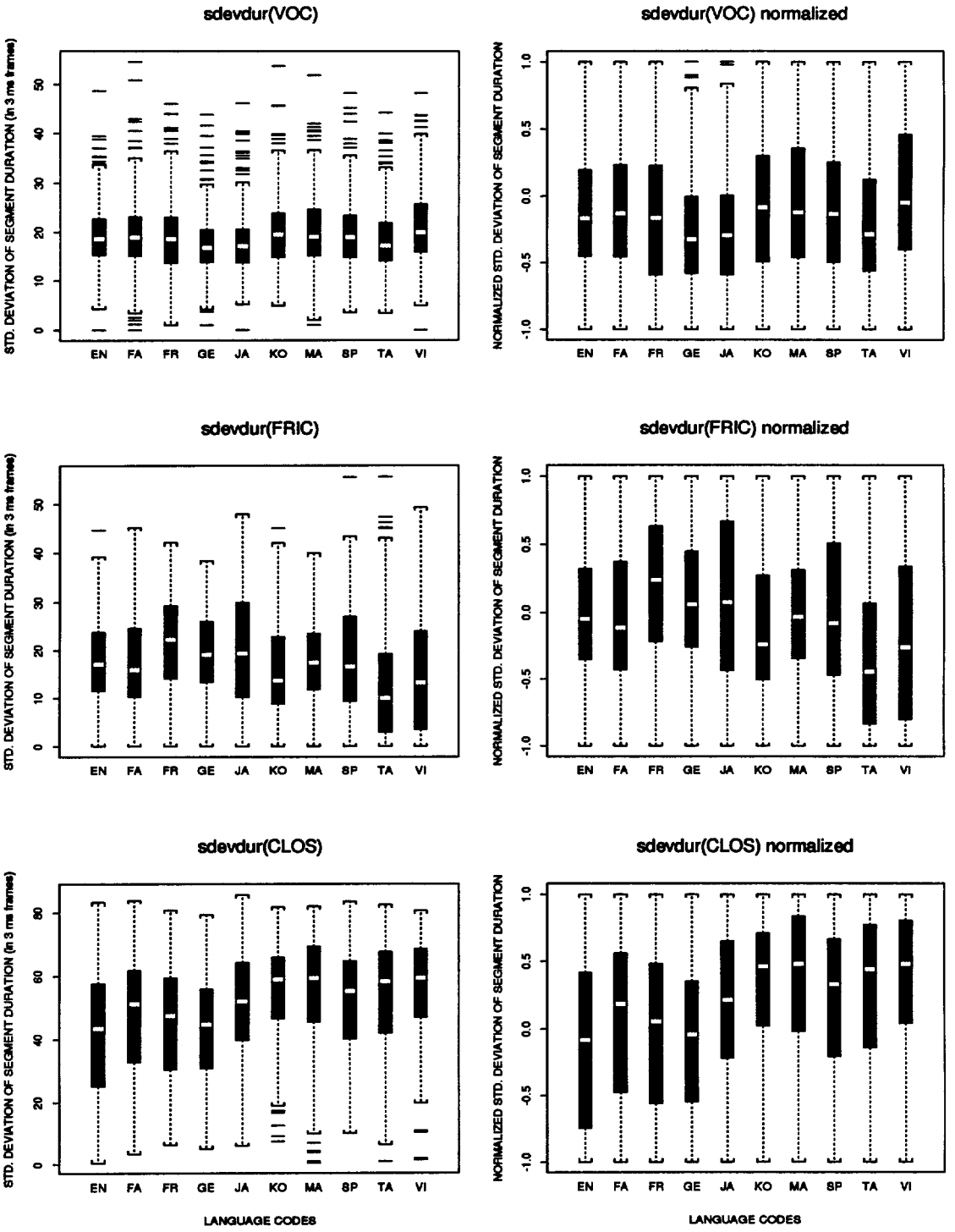

Figure C.111: Standard Deviation of Duration of VOC, FRIC and CLOS

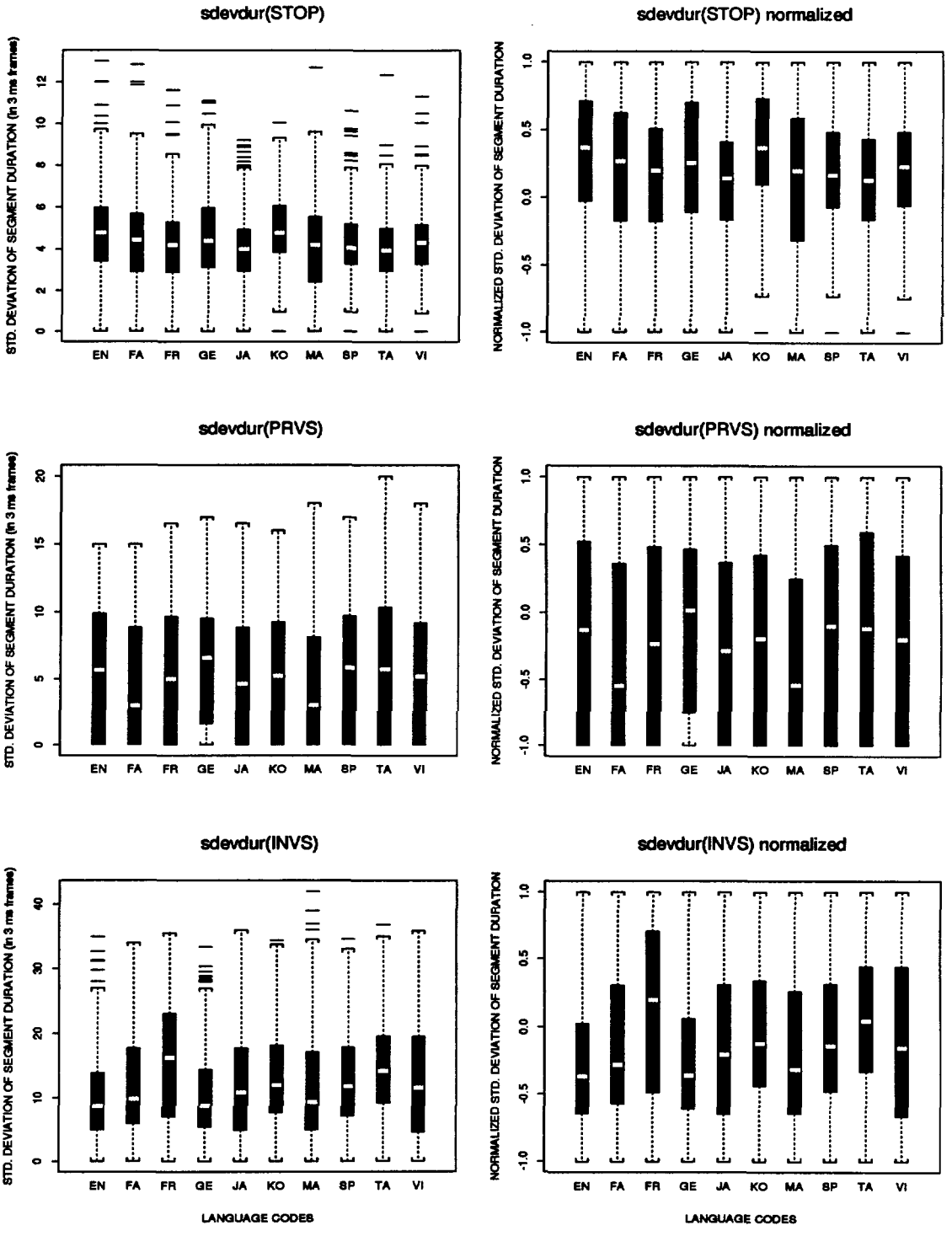

Figure C.112: Standard Deviation of Duration of STOP, PRVS and INVS

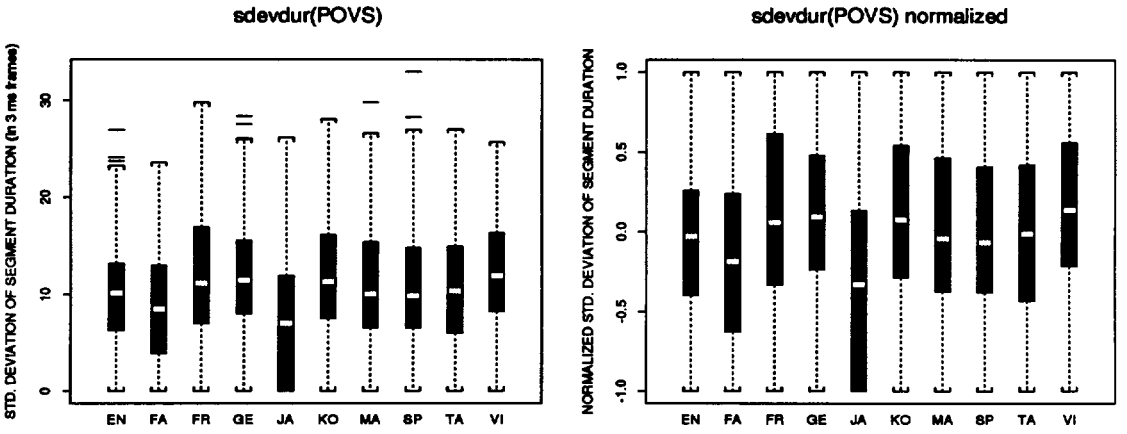

Figure C.113: Standard Deviation of Duration of POVS

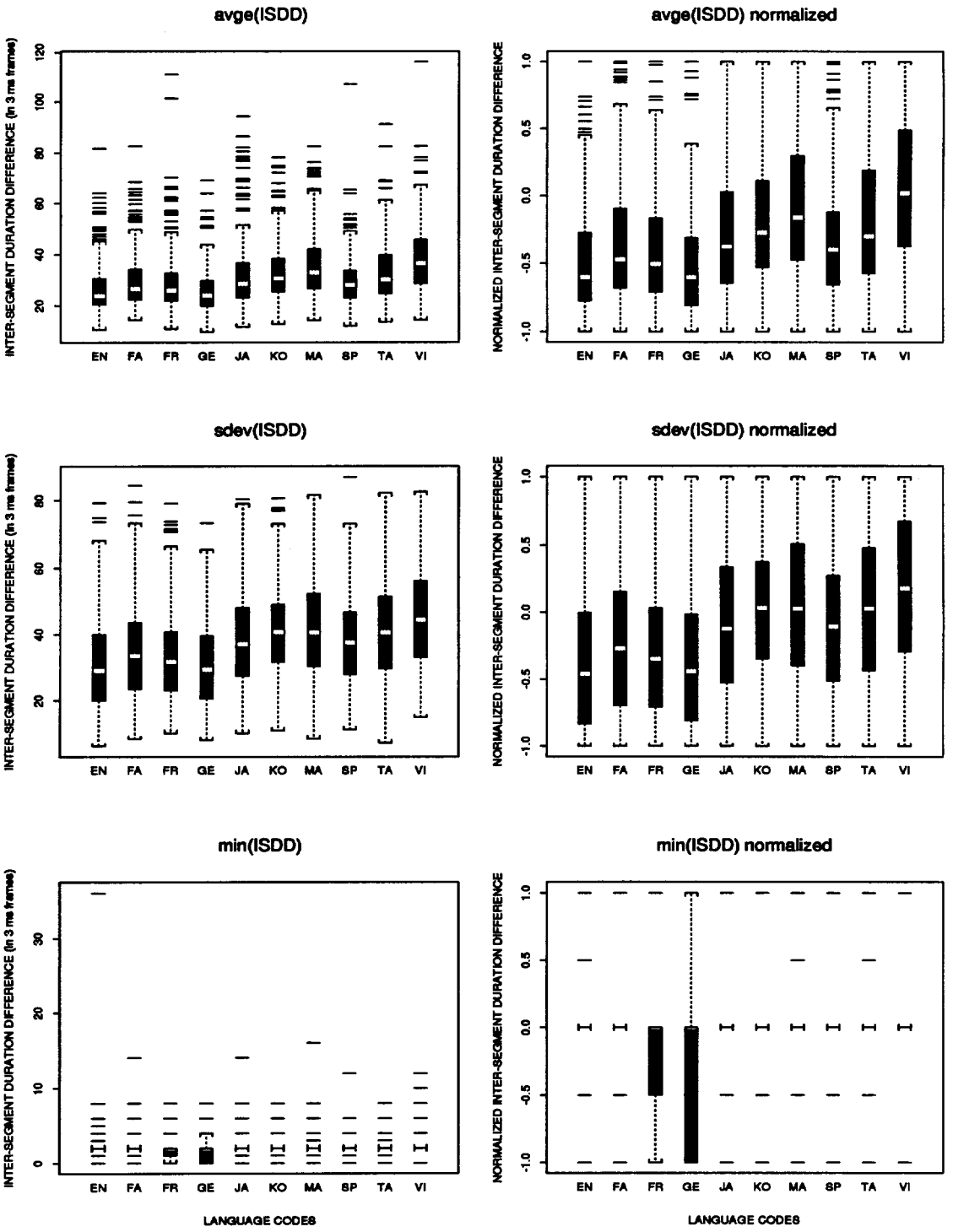

Figure C.114: Inter-segment Duration Difference Features

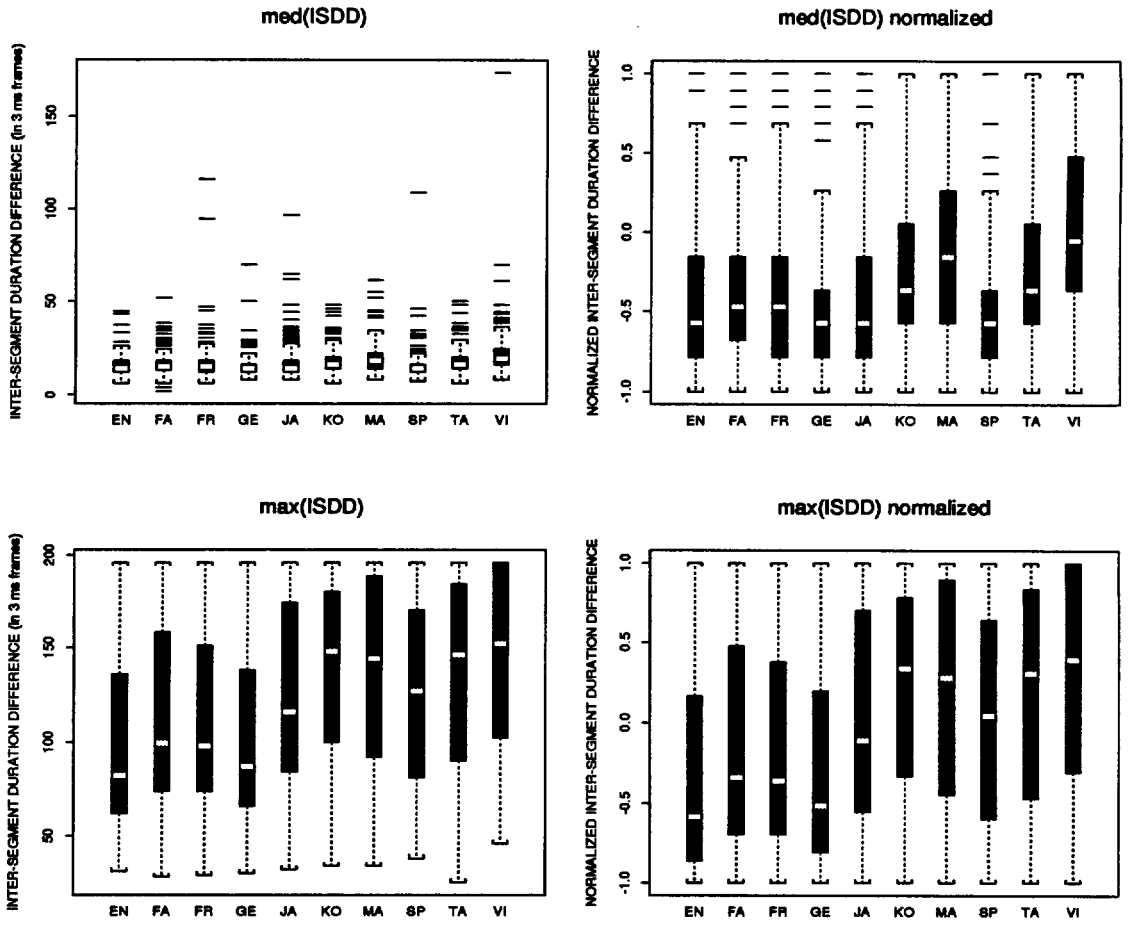

Figure C.115: Inter-segment Duration Difference Features (continued)

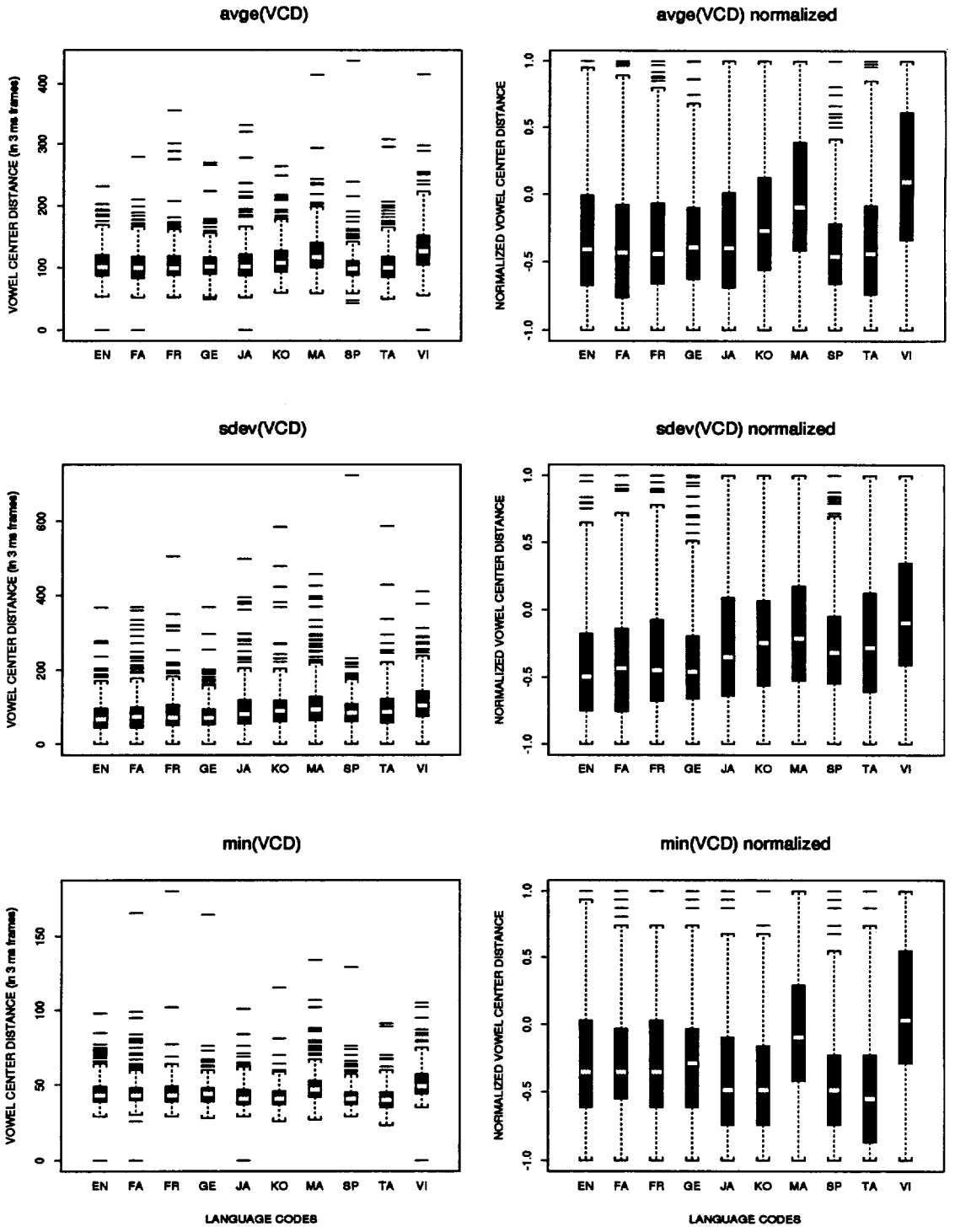

Figure C.116: Vowel Center Distance Features

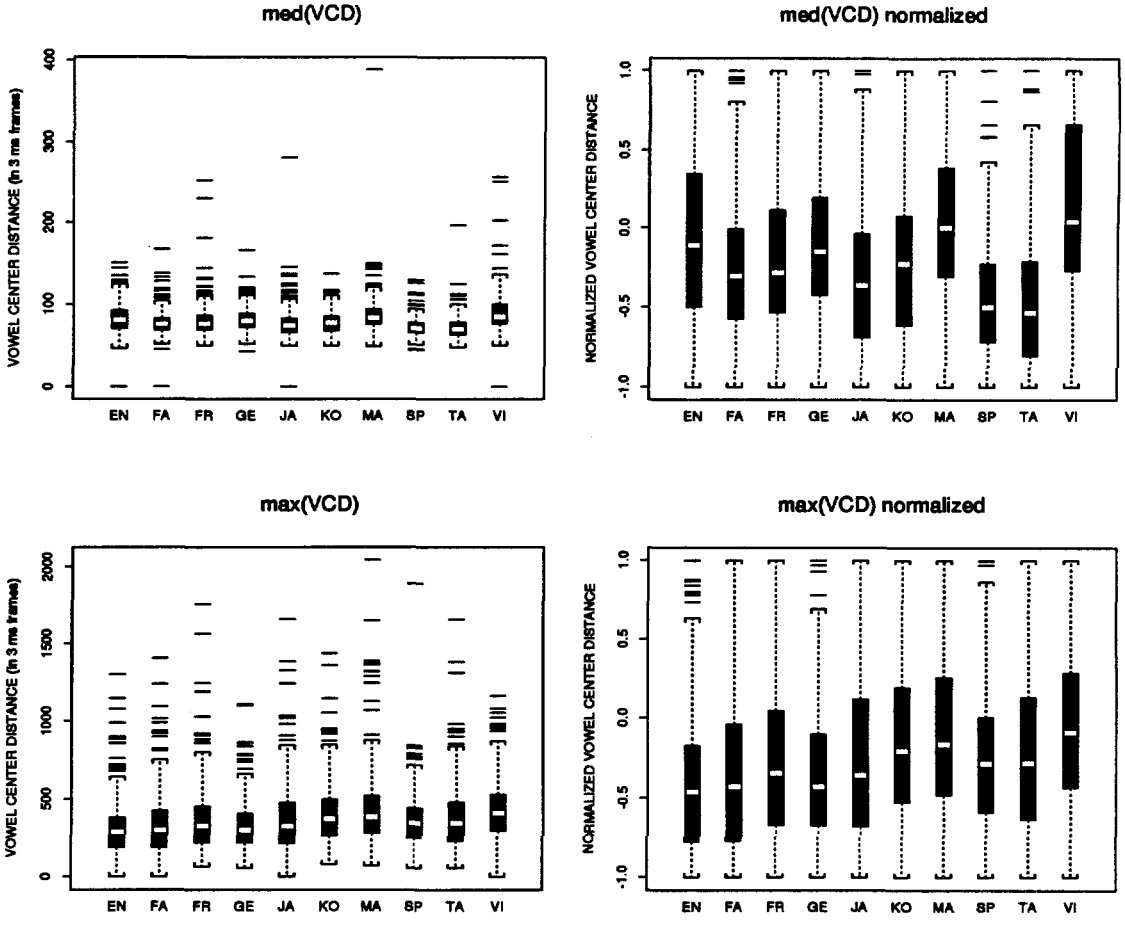

Figure C.117: Vowel Center Distance Features (continued)

Biographical Note

Yeshwant Kumar Muthusamy was born February 24, 1965 in Secunderabad, India. He graduated at the top of his class from Vidyaranya High School, Hyderabad in March 1981 with the ICSE (Indian Certificate of Secondary Education) diploma. In June 1981, he enrolled in the Hyderabad Public School at Begumpet, for his 11th and 12th grade education. He graduated in May 1983 with the ISC (Indian Secondary Certificate) diploma, majoring in Mathematics, Physics and Chemistry, and ready to join the ranks of hundreds of thousands of Indian students aspiring to become engineers.

A slew of engineering entrance examinations followed, culminating in his joining the Bachelor of Technology (B.Tech.) program in the newly created Computer Science and Engineering department at the Jawaharlal Nehru Technological University at Hyderabad in September 1983. He graduated in August 1987 at the top of his class and was awarded the University Gold Medal for his achievement. A month after receiving his B. Tech. degree, he began studies in the PhD program in the Computer Science and Engineering department at the Oregon Graduate Institute of Science & Technology (then known as the Oregon Graduate Center) in Beaverton, Oregon.

On April 8, 1988, Yeshwant attended a colloquium talk on computer speech recognition by a faculty candidate, a Speech Scientist from Carnegie Mellon University named Ronald Cole. This talk essentially altered the course of his academic career. Excited about the field, he hoped that Ron would accept a position in the department, so that he could work with him. As luck would have it, Ron did indeed accept a faculty position in the department and Yeshwant became his first student at OGI in September 1988. Projects on rule-based segmentation, vowel classification, vowel perception, English alphabet recognition (EAR) followed in quick succession before Yeshwant decided upon

the topic of his PhD dissertation: automatic language identification. Despite the advice of his friends in the speech research community, Ron gave the go-ahead for the project in March 1990. Three odd years, eons of CPU time and gigabytes of disk space later, neither advisor nor student has any misgivings about that decision.