Decision Support Systems 74 (2015) 78–87

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Financial fraud detection using vocal, linguistic and financial cues

Chandra S. Throckmorton [a], William J. Mayew [b,*], Mohan Venkatachalam [b], Leslie M. Collins [a]

[a] Electrical and Computer Engineering Department, Duke University, Durham, NC 27708, USA
[b] Fuqua School of Business, Duke University, Durham, NC 27708, USA

ARTICLE INFO

Article history:
Received 13 May 2013
Received in revised form 11 February 2015
Accepted 6 April 2015
Available online 14 April 2015

Keywords:
Automated deception detection
Financial fraud
Corporate executive speech

ABSTRACT

Corporate financial fraud has a severe negative impact on investors and the capital market in general. The current resources committed to financial fraud detection (FFD), however, are insufficient to identify all occurrences in a timely fashion. Methods for automating FFD have mainly relied on financial statistics, although some recent research has suggested that linguistic or vocal cues may also be useful indicators of deception. Tools based on financial numbers, linguistic behavior, and non-verbal vocal cues have each demonstrated the potential for detecting financial fraud. However, the performance of these tools continues to be poorer than desired, limiting their use on a stand-alone basis to help identify companies for further investigation. The hypothesis investigated in this study is that an improved tool could be developed if specific attributes from these feature categories were analyzed concurrently. Combining features across categories provided better fraud detection than was achieved by any of the feature categories alone. However, performance improvements were only observed if feature selection was used suggesting that it is important to discard non-informative features.

1. Introduction

Efficient allocation of capital is critical to stimulate economic growth, the engine that drives improvements in social welfare. The financial world has been rocked by a wave of accounting scandals, from Enron and WorldCom in the early 2000s to the collapse of Lehman Brothers in 2008 leading to a significant degradation of trust in capital markets and concomitant inefficient capital allocation decisions. In order to restore trust in the capital markets, regulators have passed a series of reforms, such as the Sarbanes Oxley Act that impose penalties on those who commit financial misdeeds. However, such penalties will be a helpful deterrent if and only if frauds are detected in a timely fashion. The current resources committed to fraud detection by the Securities and Exchange Commission (the regulatory body that oversees capital markets in the United States) are insufficient to identify all occurrences of frauds in a timely fashion. Moreover, the Center for Audit Quality, which serves the financial auditing profession, recently called for the identification of new ways to better detect financial fraud so as to restore investor confidence [31]. Thus, an automated method for accurate financial fraud detection (FFD) has the potential to deter financial misdeeds and stabilize capital markets.

Most corporate FFD methods documented in accounting and finance research have focused on financial information (see Ngai et al. [1] for a summary of this literature). This type of information includes statistics describing the company and its stock price, its financial statements, and its operating behavior, e.g. unusual reductions in employee headcount [2]. Numerous measures based on numeric financial information have been developed academically [3] and commercially [4]. Price et al. [4] examined the performance of several of these measures, and the measures all demonstrated the potential for FFD, although the commercial measure consistently outperformed the academic measures.

Very recently, several new categories of information beyond financial information have been considered for FFD. In particular, research has begun to focus on linguistic (e.g. word choices by company executives) and vocal (e.g. voice characteristics of company executives) features to assess deception. These features are based on the premise that information is contained in, and could be extracted from, what people say and how they say it [5]. Some success for FFD through linguistic information was achieved by Humpherys et al. [6] through analysis of the Management's Discussion and Analysis (MD&A) section of Form 10-Ks which public companies must file annually with the Securities and Exchange Commission (SEC). These forms are intended to provide an overview of the companies' financial and business health. By using measures of linguistic characteristics such as lexical diversity and syntactic complexity, Humpherys et al. [6] achieved up to 67% accuracy in detecting fraud. Bloomfield [16] suggested that additional insights might be gleaned from linguistic analysis of spontaneous speech of executives relative to the legally vetted text contained in MD&As. Larcker and Zakolyukina [15] undertook a linguistic text analysis of executives' conversations with financial analysts during quarterly earnings conference calls, and found that linguistic features identified deceptive discussions at better than chance levels.

* Corresponding author. Tel.: +1 19196607781.
E-mail addresses: chandra.throckmorton@duke.edu (C.S. Throckmorton), william.mayew@duke.edu (W.J. Mayew), mohan.venkatachalam@duke.edu (M. Venkatachalam), leslie.collins@duke.edu (L.M. Collins).

http://dx.doi.org/10.1016/j.dss.2015.04.006
0167-9236/© 2015 Elsevier B.V. All rights reserved.

Moving beyond linguistics, Mayew and Venkatachalam [7] studied non-verbal vocal characteristics of CEO's during earnings conference calls. They examined whether analysts and investors take into account the CEO's emotional state when they forecast future performance and if the emotional state conveyed meaningful information about future firm prospects. They found vocal based managerial affect measures contained information about a firm's financial future and that investors take this into account when determining stock prices. In a laboratory setting, Murphy [8] found that misreporters experience negative affect. Applying this intuition to corporate CEOs, Hobson et al. [9] documented that negative affect derived from non-verbal vocal features during conference calls predicts both financial fraud and the stock price reaction to the firm's disclosure of the fraud.

Collectively, tools based on financial numbers, linguistic behavior, and non-verbal vocal cues have each demonstrated the potential for detecting financial fraud. However, the performance of these tools continues to be poorer than desired, limiting their use on a stand-alone basis for identifying companies for further investigation. The hypothesis investigated in this study is that an improved FFD tool could be developed if specific attributes from across these three general feature categories were analyzed concurrently. If each of the feature categories provides independent, complementary information regarding financial fraud, then the combination of features across these categories may improve detection performance beyond what can be achieved by each feature category separately.

The potential for combined features to enhance the early identification of fraud seems plausible given the recent emergence of for-profit entities such as Business Intelligence Advisors (BIA, www.biadvisors.com). BIA hires ex-Central Intelligence Agency officers to generate reports on deception markers observed from corporate communications including earnings conference calls, and sells the reports to hedge funds and other investors interested in profiting from foreknowledge of the revelation of fraud. BIA reports are not available publicly, which prevents the testing of the results from our investigation against feature combinations executed by trained human experts. However, investigation of our hypothesis can result in initial insights on the potential benefits of feature combination via automated methods from publicly available data.

2. Material and methods

2.1. Test material

The set of 1572 public company quarterly earnings conference call audio file samples analyzed in Hobson et al. [9] were used as our sample. Earnings conference calls are ideal for our investigation because they involve corporate executives publicly discussing financial information, thereby simultaneously providing financial, linguistic and vocal cues [33].[1] The audio files contain CEO's speech during the first five minutes of the question and answer portion of the earnings conference call, enabling analysis of spontaneous executive speech in response to financial analyst inquiry. The database spans conference calls that occurred during the calendar year 2007. Uncompressed .wav files were created from an online streaming of publicly broadcast conference calls through ThomsonReuters Streetevents (www.streetevents.com) during the sample period, which were encoded in mono directly onto a computer hard disk, using Total Recorder 7.1 Professional Edition software, at 11.025 kHz sampling rate and 16 bit quantization.

To identify conference calls where executives likely deceived their investors, we followed [9] and used the Audit Analytics (www.auditanalytics.com) restatements database to find instances of financial restatements resulting from accounting irregularities. Irregularity restatements occur when financial reports are later found to be incorrect as a result of intentional managerial intervention, not clerical errors. Forty-one conference call speech samples pertain to fiscal quarters that were restated due to an irregularity. We categorized these cases as fraudulent observations and the remaining 1531 observations as non-fraudulent.

We investigate the nature of the fraud underpinning each of the 41 fraudulent observations by reviewing public filings with the SEC, in addition to legal documents and popular press articles. The misrepresented financial topics underpinning each of the frauds, and the percentage of fraudulent firm-quarter observations that pertain to each topic, are depicted in Table 1. The most frequent fraud topic in our sample pertains to revenues and sales, consistent with [3]. Frauds can stem from multiple different causes, such as a misrepresentation of both revenues and profit margins, and as such fraud topics are not mutually exclusive. Because the vast majority of frauds in our sample pertain to revenues and sales (75.6%), an analysis of whether different types of frauds are differentially predictable with numeric, linguistic and vocal cues was not possible.

2.2. Features

Our three conceptual feature categories, and the particular features we study within each category, are outlined in Table 2. Accounting risk factors are specific numeric financial features that are likely to create uncertainty for managers when reporting their firms' performance. In particular, we posit based on previous FFD research by Hobson et al. [9] and Dechow et al. [3] that large firms with poor performance, operating in highly volatile settings and higher growth expectations are more susceptible to financial restatements. Specifically, we include proxies for size, using the natural logarithm of market value of equity at the end of the fiscal quarter (lnMVE); performance, using the firm's market adjusted abnormal stock return for the preceding year (RET); operating uncertainty, using the return volatility (VOL) measured as the standard deviation of daily stock returns over the 125 trading days prior to the fiscal quarter end; and growth expectations, using the book-to-market (BM) ratio calculated as the book value of shareholders equity scaled by the market value of equity both measured at the end of the fiscal quarter. Features utilizing financial statement information are

Table 1
Fraud topics for fraudulent sample observations.

Fraud topic	% of Fraud observations pertaining to each topic
Revenue/revenue growth/sales order backlog	75.6%
Product pricing/product mix/product rollout	51.2%
Cost structure/profit margins	48.8%
Expansion/integration of acquisitions	46.3%
Selling, general and administrative	26.8%
Employee incentives/share based compensation expense/management changes	19.5%
Capital spending/production capacity/depreciation/leasing	17.1%
Strategic alliance/competition	17.1%
Inventory	14.6%
Earnings forecast/guidance	14.6%
Consolidation/restructuring/liability estimates	9.8%
Accounts receivable/allowance for doubtful accounts/securitization	9.8%
Raising capital/credit/cost of capital/return on invested capital	9.8%
Branding/advertising	9.8%
Research & development/training	7.3%
Taxes	7.3%

[1] Sample firms are traded in equity markets in the United States and sample selection details are outlined in [9]. Earnings conference calls usually occur at the end of each reporting quarter where top management make presentations to and answer questions from analysts and investors about both their firm's current reported performance as well as their future prospects. These discussions occur usually via telephone or through the internet and are often captured electronically via audio files for replay. The ThomsonReuters StreetEvents database offers restreaming of corporate earnings conference calls for a limited time following the initial broadcast.

Table 2
Description of the features.

Feature Category	Feature	Description
Accounting Risk	lnMVE	Natural logarithm of market value of equity, in millions of dollars, at the end of the fiscal quarter.
	BM	The ratio of book value of equity to market value of equity at the end of the fiscal quarter.
	VOL	Stock return volatility, measured as the standard deviation of daily stock returns over the half-year period (trading days −127 to, −2 relative to the conference call date). Variable is winsorized at the 1% and 99% level to mitigate undue outlier effects.
	RET	Current year market adjusted buy and hold stock return (i.e., firm's return minus the return to a market index). Buy and hold return is calculated for the trading days spanning the four prior fiscal quarters. Variable is winsorized at the 1% and 99% level to mitigate undue outlier effects.
Acoustic	Meanf0	Average fundamental frequency of the CEO, as measured via Praat with default system settings.
	Stdevf0	Standard deviation of the fundamental frequency of the CEO, as measured via Praat with default system settings.
	Jitter	Small-scale perturbations of the fundamental frequency of the CEO, as measured via Praat with default system settings.
	Shimmer	Variations of amplitude maxima in successive glottal cycles of the CEO, as measured via Praat with default system settings.
	Meanhnr	Mean harmonics-to-noise ratio of the CEO, where the ratio quantifies the amount of additive noise in the voice signal, as measured via Praat with default system settings.
	Stdevhnr	Standard deviation of the harmonics-to-noise ratio of the CEO, as measured via Praat with default system settings.
	Pctvoiced	Proportion of voiced speech of the CEO, as measured via Praat with default system settings.
Linguistic	LZi	The proportion of words spoken by the CEO that are first person singular pronouns multiplied by the sample median length of the transcript as in [15].
	LZwe	The proportion of words spoken by the CEO that are first person plural pronouns multiplied by the sample median length of the transcript as in [15].
	LZipron	The proportion of words spoken by the CEO that are impersonal pronouns multiplied by the sample median length of the transcript as in [15].
	LZposemone	The proportion of words spoken by the CEO that express positive emotion such as love, nice, accept etc., multiplied by the sample median length of the transcript as in [15].
	LZnegate	The proportion of words spoken by the CEO that are negative such as no, not, never, etc., multiplied by the sample median length of the transcript as in [15].
	LZcertain	The proportion of certainty words spoken by the CEO such as always, never, etc., multiplied by the sample median length of the transcript as in [15].
	LZtentat	The proportion of tentative words spoken by the CEO such as maybe, perhaps, guess, etc., multiplied by the sample median length of the transcript as in [15].
Baseline metrics	Fscore	Scaled probability of misstatement, using model developed in [3] as applied in [9].
	AR	Accounting Risk variable based on the measure developed by Audit Integrity, LLC (part of GovernanceMetrics International Ratings). Values range from 0 to 100, with higher values indicating more high risk of misstatement.
	COGDIS	Ex-Sense Pro R voice-based measure of cognitive dissonance, measured as the proportion of voice segments that register greater than 120 on the Cognition Level measure [9]. Higher values indicate more dissonance. Voice segments are approximately two-second vocal wave intervals.

extracted from the Compustat database (www.compustat.com), and features utilizing stock price performance are from the University of Chicago's Center for Research in Security Prices (www.crsp.com).

Next, we consider primitive acoustic features of voice. Nunamaker et al. [10] discuss the advancement of technology in detecting deception and allude to automated means of measuring nonverbal cues of deception. Zuckerman et al. [11] advance the theory that deceptive individuals experience internal factors such as arousal, negative affect and cognitive load that in turn leads to external displays of deception through nonverbal means. Acoustic features of the deceiver's voice can potentially offer telling cues about deception. For example, voice pitch can change due to increases in stress or nervousness from deceptive actions (Depaulo et al. [12]). Voice quality, as captured by the harmonics-to-noise ratio, is likely to be lower in deceptive individuals (Nunamaker et al. [10]). Mayew and Venkatachalam [7] document that jitter and shimmer are associated with measures of negative affect and negative stock price responses, and in turn may capture deceptive behavior. In all, we considered seven primitive acoustic features from voice in our analysis as noted in Table 2: the mean and standard deviation of the fundamental frequency of the speaker, jitter which captures pitch perturbations, shimmer which captures loudness perturbations, the mean and standard deviation of the harmonics-to-noise ratio which quantifies the 'hoarseness' of the speaker, and the proportion of voiced speech. These vocal features were extracted from the audio recordings using Praat acoustics software version 5.2.05 [13] with system default settings from the "quantifySource" GSU Praat add-on tool developed by Owren [14].

Finally, with respect to linguistic features, Vrij [18] provides an extensive review of the literature on how an individual's verbal behavior, in particular, verbal content of speech can be useful in detecting deception. Research by Knapp et al. [19] suggests that deceivers use less self-referential words; therefore, the extent to which the executives made self-referential first person singular and plural pronouns (LZi and LZwe) were determined. Similarly, the use of impersonal pronouns (LZipron) was used as a marker of deception. Research by Adams and Jarvis [20] suggests that deceivers are less likely to use positive emotion words and more likely to use negative statements; therefore, the number of positive and negative emotion words (LZPosemone and LZnegate) were measured. Finally, the proportion of tentative words (LZtentat) and words that connote certainty (LZcertain) were included to capture linguistic features that denote lack of conviction on the part of the executive (e.g., [21]).

The measurement of these linguistic features, which are prefixed by LZ to denote they follow the exact specification as in Larcker and Zakolyukina [15], was done via analyzing conference call transcripts with the Linguistic Inquiry and Word Count (LIWC) software program [30]. Conference call transcripts were obtained from ThomsonReuters StreetEvents (www.streetevents.com), and manually parsed so as to extract the question and answer portion of transcript text that matched the words spoken by the executive in the audio recording for the same fiscal quarter.[2] Eleven additional linguistic features proposed by Larcker and Zakolyukina [15] as potentially predictive of fraudulent activity were not included in our analysis. By requiring consistency between the transcript and the five minute audio files, only five minutes of speech were used to extract the linguistic features. This resulted in the majority of positive (i.e. fraudulent) and negative (i.e. non-fraudulent) examples having a value of zero for each of these eleven features, as noted in Table 3. Thus, no difference between positive and negative would be discernable for the majority of sample observations and these linguistic features were deemed *a priori* as unlikely to be informative.

In addition to primitive features in the three categories, outputs of three currently available tools were also evaluated: the academically-derived fraud score (Fscore) [3], GovernanceMetrics International's (www3.gmiratings.com, formerly Audit Integrity) commercially available Accounting Risk (AR) measure [4], and cognitive dissonance (COGDIS) as measured via analysis of each audio recording with the

[2] We use linguistic features identified in Larcker and Zakolyukina [15] because they also examine text from speech during earnings conference calls. These features are theoretically derived based on research in interpersonal deception. We do not consider features from annual report MD&A text because it is usually pre-scripted and vetted by a multitude of personnel [16] and recent research finds that linguistic features reflecting conscious deception in the MD&A do not exhibit predictive power for financial fraud [17].

Table 3
Characteristics of eleven linguistic features studied in Larcker and Zakolyukina [15] but excluded from our analysis. The prefix LZ pertains to Larcker and Zakolyukina, and the remaining feature name is the word category studied in [15].

	Mean	Standard deviation	Skewness	Excess kurtosis	Percent positive observations = Zero	Percent negative observations = Zero
LZthey	4.50	3.9	1.3	1.8	7%	10%
LZgenknlref	1.92	3.3	2.7	8.5	54%	56%
LZassent	1.60	1.6	1.1	0.9	22%	28%
LZposemoextr	4.35	3.1	1.0	1.0	10%	7%
LZanx	0.67	1.1	2.1	4.7	63%	61%
LZanger	0.64	1.0	1.9	3.8	56%	59%
LZswear	0.04	0.2	4.9	22.1	93%	96%
LZnegemoextr	1.44	1.6	1.4	1.9	29%	34%
LZhesit	0.04	0.2	4.9	22.0	98%	96%
LZshvalue	0.05	0.3	6.4	39.0	98%	98%
LZvalue	0.03	0.2	7.9	61.2	100%	98%

commercially available digital emotion analyzer Ex-Sense Pro R version 4.3.9 following [7,9]. All three measures have been shown to be related to adverse irregularity restatements resulting from intentional managerial misrepresentation of financial position in a conference call setting [9], and so we use them to provide a baseline of what can currently be achieved with this particular data set. Each of these tools forwards combinations of primitive features of either financial cues (in the case of Fscore and AR) or vocal cues from CEO speech (in the case of COGDIS) and may well consider a number of other primitive features not specifically selected for this study.[3]

2.3. Classifier and feature selection

The classifier used in this study was a generalized likelihood ratio test (GLRT) [23]. The Bayesian-based GLRT is widely used in classification applications including remote sensing, image processing, and wireless communications. It has the advantage of being computationally inexpensive while providing a more complex model of the data than the similar, commonly used naïve Bayes classifier which assumes independence between features.

The likelihood ratio is defined as

$$\lambda(x) = \frac{\int_{\theta} f(x|H_1, \theta) f(\theta) d\theta}{\int_{\theta} f(x|H_0, \theta) f(\theta) d\theta} \tag{1}$$

where $f(x|H_i, \theta)$ is the likelihood of feature values, x, for fraudulent (H_1) and non-fraudulent (H_0) companies given uncertain parameters θ that define the probability density functions (pdfs) of the features, and $f(\theta)$ is the pdf of the parameters. For the GLRT, features were assumed to be Gaussian distributed, and the mean and variance were assumed to be the maximum likelihood estimates determined from the training data. Thus, $f(\theta)$becomes a delta function located at the estimated parameters and the GLRT statistic for a feature vector x is given by:

$$\lambda_{\text{GLRT}} = \frac{\int_{\theta} f(x|H_1, \theta) \delta\left(\hat{\theta}\right) d\theta}{\int_{\theta} f(x|H_0, \theta) \delta\left(\hat{\theta}\right) d\theta} = \frac{f\left(x|H_1, \hat{\theta}\right)}{f\left(x|H_0, \hat{\theta}\right)} \tag{2}$$

Assuming that features are Gaussian-distributed is not uncommon when specific data behavior is not known *a priori*, especially in large data sets where central limit theorem holds (e.g., Duda, et al. [25]). Further, research suggests that the likelihood ratio test is fairly robust under the assumption that features are Gaussian-distributed when they are in fact from other distributions (Tantum and Collins [24]). However, some features that were originally extracted were excluded from the study due to their extreme divergence from a Gaussian distribution. These features included one linguistic feature (length of CEO speech analyzed in words) and two accounting risk features (quarterly return on assets and analyst-based unexpected earnings) due to high kurtosis and skewness indicating non-Gaussian distributions. The distributions for these three excluded features are depicted in Fig. 1.

For feature selection and testing, 10-fold cross-validation was used [25]. The data were randomly divided into ten groups or folds. One fold was held out as the test set and the remaining nine folds were used to first select a set of candidate feature combinations that could potentially result in a high level of classification and then assess these feature combinations in order to select the best performing combination. The best performing feature set was then used to test the hold-out fold. The process was repeated, using each of the 10 folds as the hold-out fold in order to determine performance across the data set. Thus all the data were used in testing, but no data were used for training and testing concurrently.

To select candidate feature combinations, the nine training folds were used in 9-fold cross-validation feature selection using an exhaustive search. While other methods of feature selection could have been used and would have converged more quickly, they are susceptible to finding local maxima in the performance space. An exhaustive search was therefore selected to give the best estimate of the highest performing combination of features. For an exhaustive search, the number of features to be included in the feature group must be selected. For this study, the number of features selected per feature combination was varied from 1 to the maximum number of features available. This resulted in N candidate feature combinations per fold (where N is the number of features); 9 N total feature combinations. Since the folds could potentially select the same feature combinations, a list of unique feature combinations was determined.

[3] The AR and COGDIS metrics are derived from commercial sources and the primitive features underpinning these metrics, as well as the algorithm to combine the underpinning features, are proprietary and unknown. See [4] and [7] for studies that utilize regression analysis to uncover potential primitive features in AR and COGDIS, respectively. Fscore is an academic measure derived as a function of financial statement line items of firms registered with the Securities and Exchange Commission in the United States, and as such we do not consider individual financial statement line items as individual features. However, in non-US countries or countries with different securities laws, investigation of whether and to what extent individual financial statement line items can assist in FFD remains an open empirical question (see for example Ravisankar et al. [22] for an analysis among Chinese firms).

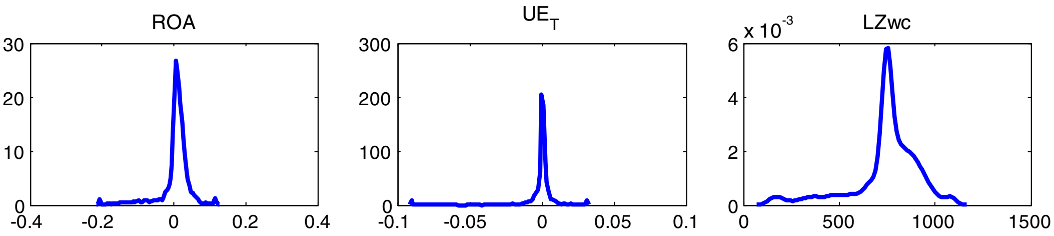

Fig. 1. Feature distributions for the three features discarded due to deviation from a Gaussian distribution. ROA is quarterly return on assets, UE_T is analyst-based unexpected earnings scaled by price per share, and LZwc is the number of words contained in the transcription of the 5 minute audio file containing CEO speech.

To evaluate each unique feature combination, 40 trials of 10-fold cross-validation using the training data were run. Performance for each trial was measured using the area under a Receiver Operating Characteristic (ROC) curve. ROC plots the probability of detecting positive occurrences, fraudulent firms in this study, versus the probability of falsely identifying negative occurrences as positive occurrences. If the performance of the classifier is equivalent to random guessing, the area under the curve (AUC) will be 0.5; if the performance is perfect, meaning all positive and no negative cases are identified as positive, then the AUC will be 1. The ROC was chosen over other common data mining performance metrics such as precision, accuracy, and f-measure due to the number of positive and negative examples in this study. These non-ROC performance metrics are sensitive to uneven distributions of observations per class, and in a case such as this study, would be driven almost entirely by the number of negative examples (Fawcett [26]).

The resulting AUCs from the 40 trials were used to generate a pdf for each feature combination. The pdfs were compared across feature combinations, and the feature combination corresponding to the best distribution was selected. The classifier was trained using the selected features from all of the training data. The trained classifier was then applied to the hold-out test data to determine performance. It should be noted that a different combination of features could be selected to test each hold out fold, and the frequency of this occurrence will be indicated in the results.

2.4. Performance evaluation

As mentioned above, the ROC was selected as the performance evaluation metric over performance metrics such as precision and accuracy due to the imbalance in the number of positive and negative examples in this data set and other large sample studies of corporate fraud [3,4,15,17]. These metrics are sensitive to observation number imbalance due to their reliance on both the number of positive and negative observations [26]:

$$accuracy = \frac{TP + TN}{P + N} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

where *TP*and *TN* refer to the number of positive and negative examples correctly identified (true positives and true negatives respectively); *FP* refers to the number of false positives (negatives incorrectly identified); and *P*and *N* refer to the total number of positives and negatives respectively. Thus, if there is a large difference between *P*and*N*, as is the case in this study, accuracy will be driven almost entirely by the class with the greater number of observations. Similarly, since *FP* is dependent on the number of negative examples available for misclassification, precision may also be driven entirely by the number of negative examples.

On the other hand, the metrics that describe a ROC, probability of detection (mathematically identical to recall) and probability of false alarm, are robust to differences in the number of observations per class because each measure relies only on responses within class:

$$Pd = \frac{TP}{P} \tag{5}$$

$$Pfa = \frac{FP}{N} \tag{6}$$

Thus neither measure is driven by the difference between *P* and *N*. Therefore, the AUC from the ROC was selected as the performance metric that would be most informative for this data set.

3. Results

3.1. Benchmark results

In Fig. 2, a ROC is plotted for the three baseline metrics: AR, COGDIS, and Fscore. The closer the lines are to the top left corner (probability of false alarm equaling zero and the probability of correct detection equaling one), the better the performance. In the legend, the AUC is listed for each baseline method. Consistent with what was observed by Price et al. [4], performance for AR was better than for Fscore. COGDIS had slightly better performance than Fscore but poorer performance than AR, which yields an AUC of 0.69. Fig. 2 simply replicates the benchmark conditions already documented in [9]. For all three metrics, however, performance is likely too poor to allow these tools to be used as standalone indicators

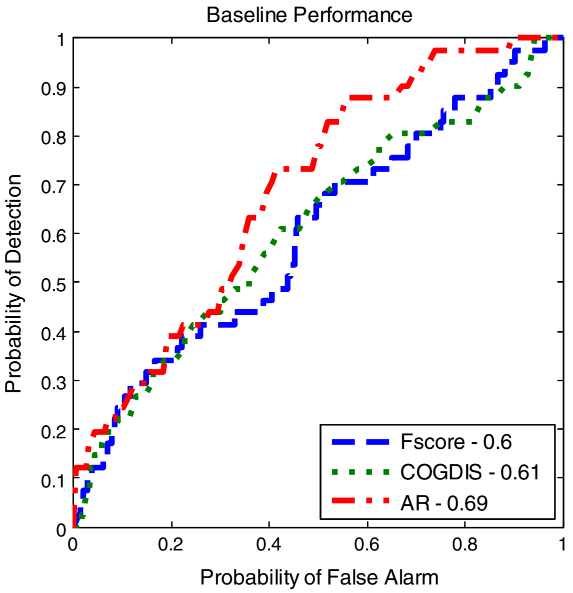

Fig. 2. Receiver operator characteristic (ROC) performance for the three baseline metrics as documented in [9]. The numbers in the legend are the Area Under Curve (AUCs) where 0.5 would indicate random guessing and 1 would indicate perfect performance. See Table 1 for feature definitions.

of probable fraudulence. Detecting roughly 70% of the accounting irregularities, while better than random chance, would still result in a large fraction of false positives.

3.2. Feature selection within feature category

The results for feature selection for each of the three feature categories are shown in Fig. 3. Each row corresponds to a feature category: accounting risk factors, acoustic features from CEO speech, and linguistic features from CEO speech. In the left column, the ROC for feature selection is plotted; and in the right column, the number of folds for which each feature was selected is shown. For accounting risk variables and acoustic features, the ROCs for the appropriate baseline metrics are included for comparison. The performance of the feature selection algorithms is similar to the performance of the baseline metrics. The performance of the accounting risk features falls between the Fscore and AR baselines. While the acoustic features provide better performance than the COGDIS baseline, their performance is similar to AR. For both the accounting risk features and the acoustic features, feature selection across the random folds was quite consistent, suggesting that the selected features are consistently informative regardless of data partitioning.

For the linguistic feature category, however, only one feature was chosen for more than half of the folds, suggesting that feature selection had difficulty finding a combination of features that classified the data well. This is further demonstrated by the ROC. The selected linguistic features provided no reliable indication of fraudulent firms (AUC = 0.5). This result differs from Larcker and Zakolyukina [15], who found linguistic features to be reliable fraud indicators. We discuss in Section 4.1 potential explanations for the poor predictive ability of linguistic features in our sample relative to [15].

3.3. Feature selection across feature categories

To assess the potential benefit of combining features across feature categories, the features selected from the training data for each feature category were combined and used to test the hold-out fold. Only the selected acoustic and accounting risk features were used since the linguistic statistics provided no predictive power in this study. The ROC for the combined features is compared to the three baseline metrics in Fig. 4a. Combining features across categories had a positive impact on performance, providing better performance than the best baseline metric (AR) at most operating points on the ROC, providing a 15% reduction in false alarms at 90% detection. However, the advantage of combining features across feature categories is dependent on the selection of the features for inclusion. The ROC that results from using all of the acoustic and accounting risk features is plotted in Fig. 4b. If all of the features are used, performance drops to a level below AR. These results suggest that while using features across categories rather than a single category of features can provide an increase in performance, it is also important to

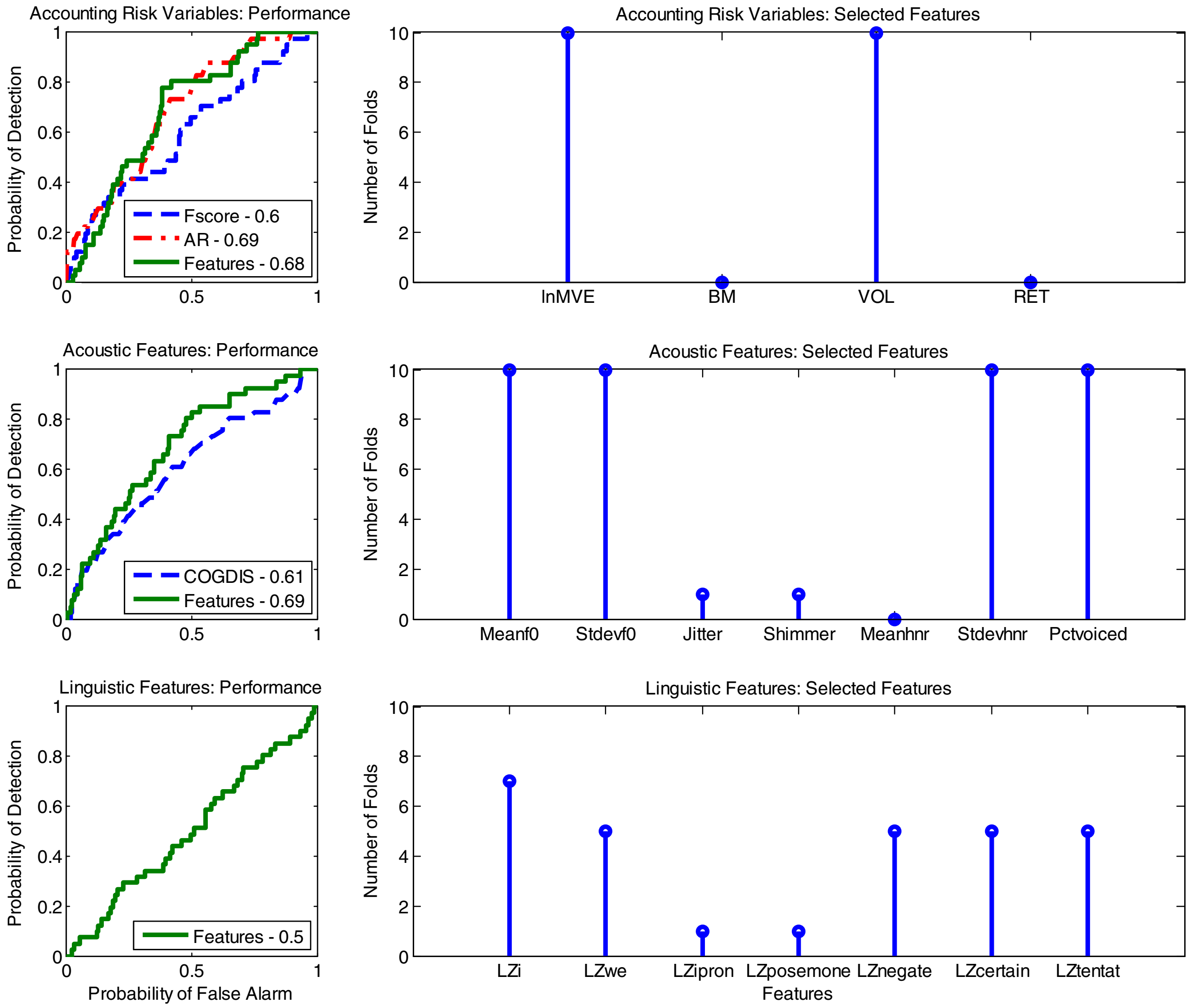

Fig. 3. Performance and selected features for each feature category. Each row is a different feature category. The left column is the ROC for feature selection with ROC's of the corresponding baseline metrics where applicable. The right column is the number of folds for which each feature was selected. See Table 1 for feature definitions.

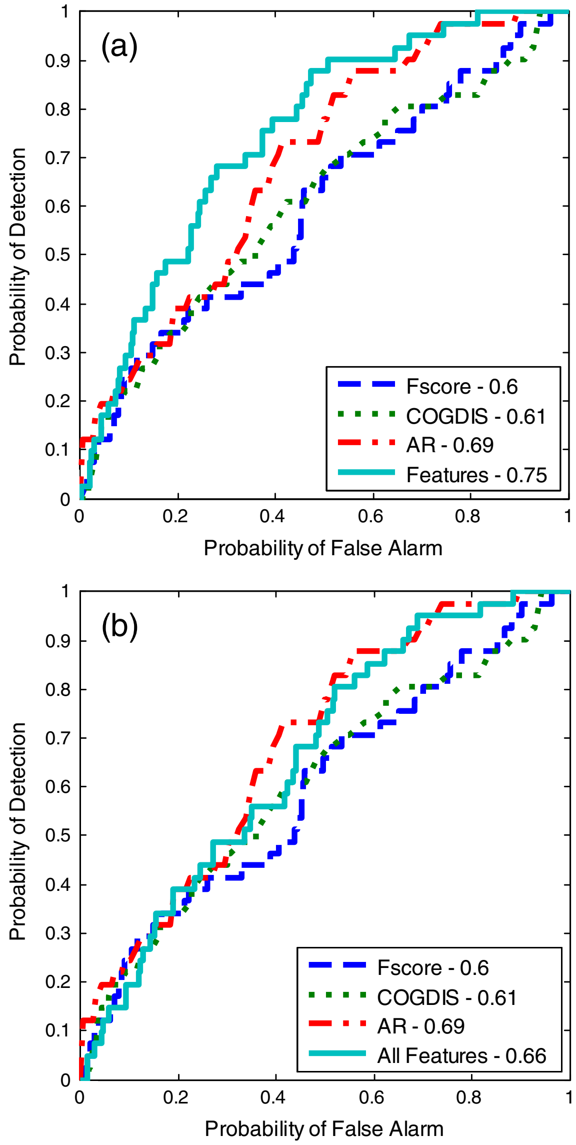

Fig. 4. Performance using features across all feature categories (a) with and (b) without feature selection. (a) The features were those selected by the training data for the acoustic and accounting risk feature categories. (b) Performance using all acoustic, accounting risk, and linguistic features. See Table 1 for feature definitions.

perform feature selection to ensure that the informative features are used.

Given the importance of using informative features, a final analysis displayed in Fig. 5 was conducted for which the outcomes of the three baseline metrics were included with the selected features. These three metrics have been previously shown to individually provide some of the highest levels of FFD in large archival samples and therefore can be considered highly informative features [4,9]. Fig. 5 reveals that inclusion of these features results in a further performance gain over combining the features extracted in this study. A further gain is achieved by adding only the most predictive of the baseline metrics (AR). These results support the conclusion that performance gains are dependent on excluding less informative features.

3.4. Classifier sensitivity and classifier data partition sensitivity

The GLRT was chosen as a classifier due to its lower computational complexity which made feasible the use of an exhaustive search for feature selection. The use of an exhaustive search for feature selection precluded the use of such common data mining classifiers as vector machines or decision trees due to computational complexity; however, other classifiers with lower computational complexity are routinely used in data mining, and it is possible that other classifiers might outperform the GLRT. Three additional classifiers were considered: logistic regression, naïve Bayes, and K-nearest neighbors (KNN) [25] and their performance is compared with the GLRT in Table 4. The top two rows demonstrate the benefit of discarding non-informative features. When the linguistic features are discarded, performance improves across all the classifiers. The success of the classifier-based feature selection (Combined Features row) was mixed with the logistic regression and KNN classifiers receiving little to no benefit from further feature selection.

The GLRT tends to provide the best performance among this group of classifiers. The logistic regression and naïve Bayes classifiers both assume independence between features which is a simpler model than that provided by the GLRT which estimates the covariance between the features. The KNN classifier is non-parametric and thus does not assume a model; however, since its decision is based on the number of neighbors, it can be negatively impacted by an uneven distribution of the number of observations per class. These characteristics of the other classifiers may have limited their ability to make effective use of the acoustic features. While all the classifiers had comparable performance with the accounting risk features, the GLRT had noticeably higher performance with the acoustic features than the other classifiers. By failing to make effective use of the acoustic features, the classifiers

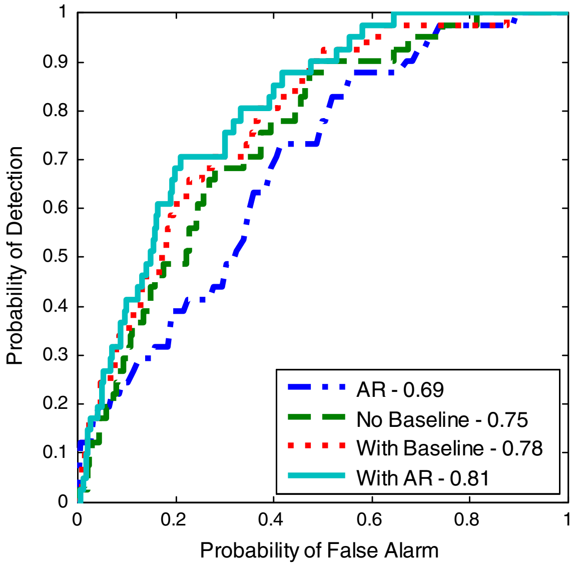

Fig. 5. Performance for combined features under three conditions: without the baseline metrics included, with all three baseline metrics included, and with only AR added to the feature set. The best baseline metric, AR, is included for comparison. See Table 1 for feature definitions.

Table 4
AUC performance of feature selection with additional classifiers.

	GLRT	Naïve Bayes	Logistic Regression	KNN
All features, all types	0.68	0.63	0.58	0.65
All features, without linguistic	0.72	0.64	0.61	0.67
Accounting risk features	0.68	0.68	0.65	0.74
Acoustic features	0.69	0.52	0.45	0.59
Linguistic features	0.50	0.54	0.53	0.45
Combined features	0.75	0.68	0.62	0.65
Combined features with all three baseline metrics	0.78	0.76	0.74	0.72
Combined features with only AR baseline metric	0.81	0.76	0.72	0.75

other than the GLRT may not have been as able to benefit from combining feature types (see Combined Features row); although all the classifiers did benefit from the addition of the presumably highly informative baseline features. Thus, a classifier that provides a more complex model of the data may be necessary to take advantage of the addition of some feature types. It is possible that performance would be further improved by selecting a more computationally complex classifier with a more detailed model for the data; however, the trade-off with using a different method of feature selection would need to be assessed.

While an imbalance between the numbers of observations per class should not theoretically impact classifier training, it may be possible that changes in data partitioning would result in large changes in performance. This would be especially true if the positive samples differ substantially from one another such that data used for training differs substantially from test data. To verify that this did not occur, pdfs of AUC for the five feature groups were generated and plotted in Fig. 6. Each pdf is the result of 100 trials of 10-fold cross-validation applied to the entire data set. No new feature selection was conducted; the accounting risk features used were lnMVE and VOL and the acoustic features were Meanf0, Stdevf0, Stdevhnr, and Pctvoiced, based on results from Fig. 4. The ranking of the different conditions matches the observed performance in Figs. 4 and 5, suggesting that the results are robust despite the small sample size.

3.5. False alarms at various operating points

The ROC plots all possible levels of detection and false alarm; however, in a final system, the user must select a desired level of performance, termed the “operating point”. Examples of operating points are shown in Fig. 7 for the best and worst performing feature sets for the GLRT. For example, the user might prefer that all positive cases be detected (probability of detection = 1). This would result in probabilities of false alarm equal to 0.65 for the combined features and 0.76 for the accounting risk features which is the equivalent of 996 and 1164 false alarms respectively for this data set. This number of false alarms is likely prohibitive to using these algorithms. However, it is possible that a number of missed detections would be acceptable in the system, especially if it can be assumed that the positive examples that are hardest to detect are less likely to be the most egregious fraudulent examples. Selecting a level of detection lower than 100% may then result in an acceptable number of false alarms.

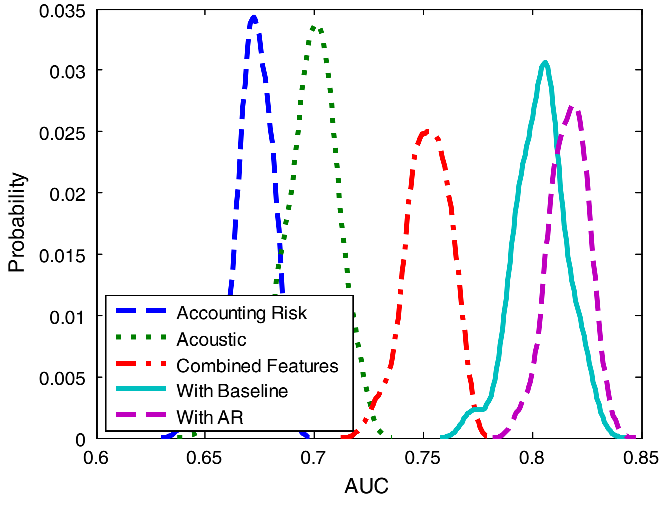

Fig. 6. Probability density functions of AUC for five different combinations of features using 100 trials of 10-fold cross-validation across all of the data. Accounting risk features were lnMVE and VOL; acoustic features were Meanf0, Stdevf0, Stdevhnr, and Pctvoiced; combined features were the 6 accounting risk and acoustic features; combined features with baseline were the 6 combined features and Fscore, COGDIS, and AR; and combined features with AR were the 6 combined features and AR only. See Table 1 for feature definitions.

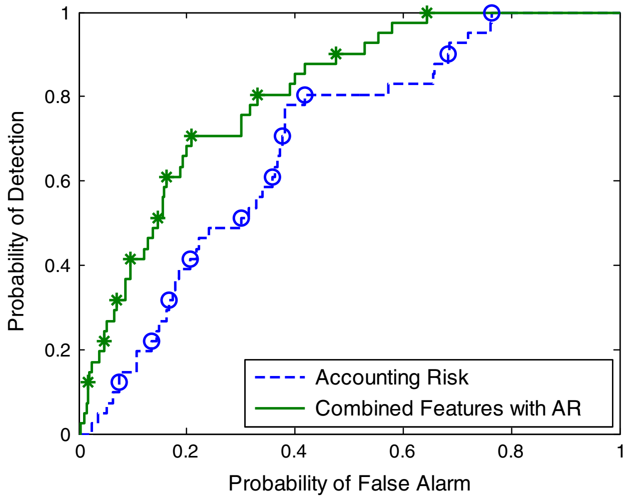

Fig. 7. Example operating points shown on two ROCs for the best performing and worst performing feature sets for the GLRT. The operating points are located at probabilities of detection = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.

With this in mind, the number of false alarms was calculated for the five feature groups for several operating points, as displayed in Table 5. By setting the operating point to 20% detection of fraudulent examples, the number of false alarms drops to less than 100 non-fraudulent examples for most of the feature sets. If it can be assumed that this detection level would be the equivalent of detecting the most egregious 20% of the fraudulent cases, then this operating point might be acceptable for improving the health of the financial market despite missing the detection of the less egregious 80% of fraudulent cases. However, it remains to be tested whether greater classifier confidence in fraudulence corresponds to the degree of fraudulence such that operating at a lower point of probability of detection would be acceptable.

4. Discussion

Recent research on financial fraud detection has expanded the categories of features that may prove useful beyond accounting risk factors derived from firm financial characteristics; in particular, the role of verbal and nonverbal features. However, these studies tend to consider the potential for these new feature categories in isolation. The hypothesis of this study was that using specific features across the feature categories might provide improvements in accuracy and error rate that are not observed from isolated feature categories. One challenge to testing this hypothesis was the extraction of suitable features for comparison to the commercial AR and COGDIS metrics, since the actual features used in these metrics and the algorithm to combine the features are unknown. The plots in Fig. 3 suggest that the features used in this paper, when considered in isolation, provide similar performance to the commercial metrics. Thus, any observed performance improvements over the baseline algorithms are likely due to selecting

Table 5
Number of false alarms for a given probability of detection.

	90%	80%	70%	60%	50%	40%	30%	20%	10%
Accounting risk features	1047	640	578	549	370	284	250	165	97
Acoustic features	998	730	625	532	378	277	194	92	42
Combined features	775	680	516	373	265	203	142	88	35
Combined features with all three baseline metrics	724	560	417	297	254	171	113	60	18
Combined features with only AR baseline metric	729	507	321	249	211	145	101	59	26

and using features across categories rather than defining features that inherently have more predictive power.

Using features across categories demonstrates the potential for improvement in detection and error rate (Figs. 4a and 5). However, discarding non-discriminative features appears to be important for achieving these improvements. In Fig. 4b, all performance improvement was lost when no feature selection was conducted. Thus, while utilizing features across categories can result in substantial performance improvements; these results suggest that it is important to include only informative features. This is emphasized in the difference between the performance improvements seen in Figs. 4a and 5. The commercial algorithms can be assumed to be the result of significant optimization efforts, rendering the assumption that their outputs would be highly relevant to fraud and deception detection. Using these features resulted in an additional increase in performance, emphasizing the importance of the quality of features to final performance.

4.1. Lack of predictive power of select features

One surprising result was that the linguistic features had no predictive power in this study. This differs from the recent analysis by Larcker and Zakolyukina [15], which documented that linguistic features predicted fraud at levels 6–16% above chance in a conference call setting. Sample differences may account for this difference for two reasons. First, Larcker and Zakolyukina [15] study over 17,000 conference calls, while we study only 1572 calls. Our sample is much smaller because we study vocal features in addition to linguistic features, and obtaining vocal features is much more difficult and costly than obtaining linguistic features [33]. Perhaps we do not have enough statistical power in our sample to observe the linguistic effects documented in [15]. Second, for consistency, in the present study, both the acoustic and the linguistic features were extracted from the same five minute CEO-Q&A excerpt, with our median transcript containing 751 words. In contrast, Larcker and Zakolyukina [15] relied on CEO speech in the entire conference call, which resulted in a median transcript length of 2902 words, roughly four times as large as the speech samples analyzed here.

It seems plausible that the amount of text analyzed will have an impact on the predictive power of linguistic statistics. For example, a single sentence is unlikely to provide much information as almost any "bag of words" frequency count will result in a value of zero. In this study eleven out of 19 of the proposed features in [15] were discarded due to the high proportion of positive and negative examples that had feature values equal to zero as summarized in Table 3. It is possible that to be predictive, a greater proportion of these features must be included, therefore suggesting that longer speech segments are required for predictive power. We note, however, that we observe similar base proportions for linguistic features in Table 3 as are documented in [15], suggesting the possibility that differences in text length may not fully explain why we observe no predictive power for linguistic features.

Further study is ultimately required to determine the relationship between the predictive power of linguistic features and the amount of text available for analysis. As [15] represented the first exploratory study of deception detection in an earnings conference call setting, much opportunity exists for a further investigation of generalizability as well as ways to increase the power of linguistic analysis [32]. Future research might also consider whether an expanded set of linguistic features derived from data-mined words rather than theoretically derived words from the deception literature can enhance predictive power [17].

Another unexpected result from the feature selection was that features that were hypothesized or have been demonstrated to be indicative of deception in isolation were not selected for the final algorithm. For example, the harmonics-to-noise ratio (Meanhnr) has been demonstrated to be indicative of deception (e.g., [27]) but was not selected. It is possible that the non-selected features, while informative in isolation, were redundant in terms of discrimination with other selected features.

4.2. Alternative classifier and feature selection methods for enhanced performance

In this study, a classifier and a method of feature selection had to be selected. The exhaustive search was selected due to its robustness to local maxima in the performance space; however, this feature selection method limited the selection of classifiers to those with low computational complexity in order to make feature selection feasible. However, the results shown in Table 4 suggest that performance might be improved by using classifiers that model more complex relationships between the features, especially for the acoustic features. The GLRT provided better performance than the naïve Bayes and logistic regression classifiers, and one of the most significant differences between these classifiers is that the GLRT does not assume independence between features. While it is possible that the observed positive outcome is specific to using a GLRT as the classifier, it is more likely that the outcome is a result of relying on a classifier that better models the data. Other classifiers such as decision trees and vector machines are routinely used in data mining, and it is possible that classifiers from either of these categories might improve performance beyond that achieved by the GLRT. However, the trade-off would be that an exhaustive feature selection method could not be used. A more rapid approach to feature selection would be to use a filter method which applies an algorithm directly to the feature values and based on some outcome decides the utility of the feature (Saeys et al. [28]; Ravisankar et al. [22]). The disadvantage of this approach is that the "good" outcome from the filter that leads to a feature being selected may not indicate high performance with the chosen classifier. Alternatively, other wrapper methods, which assess features using the desired classifier, such as a genetic algorithm or sequential forward search may provide faster convergence and ideally will converge on the same outcome as an exhaustive search. Further investigation will be required to determine if there is an advantage to using a more computationally complex classifier and whether other feature selection methods impact performance.

4.3. Differential costs of false positives and false negatives by stakeholder

The utility of the FFD algorithm presented here may depend on the stakeholder interested in detecting financial fraud. Different stakeholders may use different likelihood thresholds for determining the severity of the fraud depending on the cost of false positives and false negatives. For example, an auditor may worry about false positives to avoid unnecessary costly audit procedures whereas a hedge fund manager may worry less about false positives because they follow a portfolio approach that minimizes such costs. In the case of the former, it may be possible to reduce the number of false alarms by searching for the most egregious cases of fraud rather than attempting to detect all cases of fraud. The results from Table 5 suggest that if threshold is set such that only the 10% of fraudulent cases with the highest likelihood of being fraudulent are detected, then the number of false alarms drops well below 50. However, further investigation will be required to determine whether highest likelihood from the classifier corresponds to most egregious cases in terms of the impact of the fraud.

5. Conclusion

This study examined whether an improved FFD tool could be developed by combining specific features derived from financial information and corporate executive speech, rather than examining features in isolation. The hypothesis that the feature categories from numeric financial data, linguistics and non-verbal vocal cues provide complementary information for FFD is supported by the results in this paper. By identifying which features are informative for FFD and which are not, this study answers the call by Ngai et al. [1] for research that can guide practitioner (i.e. regulators, financial analysts, auditors and investors) implementation of data mining techniques for FFD. However, these results have been tested on only one data set, albeit with careful cross-validation, and further validation is needed. These results also suggest that performance improvements depend on the quality of the features. Research defining and optimizing features is likely to continue to have a significant impact on the development of tools for fraud detection, given both the growth in available features for analysis and the propensity of managers to evolve so as to evade detection [29].

References

[1] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: A classification framework and academic review of literature, Decision Support Systems 50 (2011) 559–569.

[2] J.F. Brazel, K.L. Jones, M.F. Zimbelman, Using nonfinancial measures to assess fraud risk, Journal of Accounting Research 5047 (2009) 1135–1166.

[3] P.M. Dechow, W. Ge, C.R. Larson, R.G. Sloan, Predicting material accounting misstatements, Contemporary Accounting Research 28 (2011) 17–82.

[4] R.A. Price III, N.Y. Sharp, D.A. Wood, Detecting and predicting accounting irregularities: A comparison of commercial and academic risk measures, Accounting Horizons 25 (2011) 755–780.

[5] C. Caffi, R.W. Janney, Toward a pragmatics of emotive communication, Journal of Pragmatics 22 (1994) 325–373.

[6] S.L. Humpherys, K.C. Moffitt, M.B. Burns, J.K. Burgoon, W.F. Felix, Identification of fraudulent financial statements using linguistic credibility analysis, Decision Support Systems 50 (2011) 585–594.

[7] W.J. Mayew, M. Venkatachalam, The power of voice: Managerial affective states and future firm performance, Journal of Finance 67 (2012) 1–43.

[8] P.R. Murphy, Attitude, Machiavellianism and the rationalization of misreporting, Accounting, Organizations and Society 37 (2012) 242–259.

[9] J.L. Hobson, W.J. Mayew, M. Venkatachalam, Analyzing speech to detect financial misreporting, Journal of Accounting Research 50 (2012) 349–392.

[10] J.F. Nunamaker, D.C. Derrick, A.C. Elkins, et al., Embodied conversational agent-based kiosk for automated interviewing, Journal of Management Information Systems 28 (1) (2011) 17–48.

[11] M. Zuckerman, B.M. DePaulo, R. Rosenthal, Verbal and nonverbal communication of deception, Advances in Experimental Social Psychology 14 (1) (1981) 1–59.

[12] B.M. DePaulo, J.J. Lindsay, B.E. Malone, et al., Cues to deception, Psychological Bulletin 129 (1) (Jan, 2003) 74–118.

[13] P. Boersma, D. Weenink, Praat: Doing phonetics by computerAvailable: www.praat.org.

[14] M.J. Owren, GSU Praat Tools: Scripts for modifying and analyzing sounds using Praat acoustics software, Behavior Research Methods 40 (2008) 822–829.

[15] D.F. Larcker, A.A. Zakolyukina, Detecting deceptive discussions in conference calls, Journal of Accounting Research 50 (2012) 495–540.

[16] R. Bloomfield, Discussion of annual report readability, current earnings and earnings persistence, Journal of Accounting and Economics 45 (2008) 248–252.

[17] L. Purda and D. Skillicorn. "Accounting Variables, Deception and a Bag of Words: Assessing the Tools of Fraud Detection," *Contemporary Accounting Research*, 2014, Online Early DOI: 10.1111/1911-3846.12089

[18] A. Vrij, Detecting lies and deceit: Pitfalls and opportunities, 2nd ed. Wiley, 2008.

[19] M.L. Knapp, R.P. Hart, H.S. Dennis, An exploration of deception as a communication construct, Human Communication Research 1 (1974) 15–29.

[20] S.H. Adams, J.P. Jarvis, Indicators of veracity and deception: An analysis of written statements made to police, Speech, Language, and the Law 13 (2006) 1–22.

[21] M.L. Newman, J.W. Pennebaker, D.S. Berry, J.M. Richards, Lying words: Predicting deception from linguistic styles, Personality and Social Psychology Bulletin 29 (2003) 665–675.

[22] P. Ravisankar, V. Ravi, G. Raghava Rao, I. Bose, Detection of financial statement fraud and feature selection using data mining techniques, Decision Support Systems 50 (2011) 491–500.

[23] S.M. Kay, Fundamentals *of Statistical Signal Processing*, 1st ed., Detection Theory, vol. II, Prentice Hall, 1998.

[24] S.L. Tantum, L.M. Collins, Detection and classification of landmine-like targets in a non-Gaussian noise environment, Proceedings of SPIE 4038 (2000) 900–909.

[25] R.O. Duda, P.E. Hart, D.G. Stork, Pattern classification, 2nd ed. John Wiley & Sons, Inc, New York, New York, 2001.

[26] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861–874.

[27] J.F. Nunamaker, J.K. Burgoon, N.W. Twyman, J.G. Proudfoot, R. Schuetzler, J.S. Giboney, Establishing a foundation for automated human credibility screening, IEEE International Conference on Intelligence and Security Informatics, 2012.

[28] Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517.

[29] W. Zhou, G. Kapoor, Detecting evolutionary financial statement fraud, Decision Support Systems 50 (2011) 570–575.

[30] J.W. Pennebaker, M.E. Francis, R.J. Booth, Linguistic Inquiry and Word Count: LIWC 2001, Lawrence Erlbaum Associates, Mahway, 2001.

[31] Center for Audit Quality, Deterring and detecting financial reporting fraud – a platform for action, www.thecaq.org2010.

[32] R. Bloomfield, Discussion of detecting deceptive discussions in conference calls,, Journal of Accounting Research 50 (2012) 541–551.

[33] W.J. Mayew, M. Venkatachalam, Speech analysis in financial markets, Foundations and Trends in Accounting 7 (2012) 73–130.

Chandra S. Throckmorton is a Senior Research Scientist in the Department of Electrical and Computer Engineering at Duke University. She has published in such journals as the Journal of the Acoustic Society of America, IEEE Transactions on Aerospace and Electronic Systems, and IEEE Transactions on Biomedical Engineering Hearing Research.

William J. Mayew is an Associate Professor of Accounting at Duke University's Fuqua School of Business. He has published in such journals as the Journal of Finance, Journal of Accounting Research and The Accounting Review.

Mohan Venkatachalam is a Professor of Accounting at Duke University's Fuqua School of Business. He has published in such journals as the Journal of Finance, Journal of Accounting Research and Journal of Accounting and Economics.

Leslie M. Collins is a Professor of Electrical and Computer Engineering as well as a Professor of Biomedical Engineering at Duke University. She is also a senior member of the IEEE, publishing in such journals as the Journal of the Acoustic Society of America and IEEE Transactions on Signal Processing.