COMBINING PROSODIC LEXICAL AND CEPSTRAL SYSTEMS FOR DECEPTIVE SPEECH DETECTION

Martin Graciarena[*] *Elizabeth Shriberg*[†*] *Andreas Stolcke*[†*]
Frank Enos[**] *Julia Hirschberg*[**] *Sachin Kajarekar*[*]

[*]SRI International, USA
[†]International Computer Science Institute, USA
[**]Columbia University, USA
martin@speech.sri.com

ABSTRACT

We report on machine learning experiments to distinguish deceptive from nondeceptive speech in the Columbia-SRI-Colorado (CSC) corpus. Specifically, we propose a system combination approach using different models and features for deception detection. Scores from an SVM system based on prosodic/lexical features are combined with scores from a Gaussian mixture model system based on acoustic features, resulting in improved accuracy over the individual systems. Finally, we compare results from the prosodic-only SVM system using features derived either from recognized words or from human transcriptions.

1. INTRODUCTION

The automatic detection of deceptive speech is of particular interest to law enforcement and other government agencies, for example, in evaluating reports from informants at embassies and consulates throughout the world, in identifying potential deception in border crossings, and as an antifraud tool.

Most studies in the literature on deceptive behavior have involved human perception evaluations or descriptive analyses of facial, gestural, and biometric data. Significant research has been done in the psychology of deceptive behavior, where the main focus has been on identifying visual cues (body and facial gestures) through laboratory experiments (see [2] for a literature review in this area).

A few studies have included audio analysis: Ekman et al. [4] found a significant increase in pitch for deceptive speech over truthful speech. Streeter et al. [11] reported similar results, with stronger findings for more highly motivated subjects. De-Paulo et al., in their meta-study of previous research findings in deception [2], reported significant effects for increased pitch and vocal tension in their overall examination of evidence of subject 'tenseness' during deception. There is also some literature by members of law enforcement agencies and the military identifying auditory and lexical cues to deception. The most widely cited sources include response latency, filled pauses, coherence of discourse, passive voice, and use of contractions [1, 8]. Voice stress analysis procedures attempt to rely on low-level indicators of stress as indirect indicators of deception [5]. However, despite some evidence from the research community and belief among practitioners, there has been little work on the *automatic* identification (by machine) of deceptive speech from such acoustic, prosodic, and lexical cues.

Recently, a corpus-based machine learning approach combining lexical, prosodic, and speaker-dependent features for distinguishing deceptive from nondeceptive speech was presented [6]. In that reference the Columbia-SRI-Colorado (CSC) corpus was introduced. Our work also uses that corpus but explores (especially acoustic) cues not previously applied to this task, and focuses on system combination and issues arising from automatic speech recognition.

In this paper we first describe the CSC corpus. In Section 3 we describe the features and classifiers from each individual system, and the system combiner. In Section 4 we describe the experiments performed. In Section 5 we present the conclusions followed by the references.

2. THE CSC CORPUS

One of the primary obstacles to research in automatic deception detection from speech is the lack of a cleanly recorded corpus of deceptive and nondeceptive speech for training and testing. Existing corpora are difficult to analyze because of poor recording conditions. While early studies were better able to utilize scenarios with 'high stakes' deception in the laboratory (in which subjects could be motivated by fear or shame) [7], more recent studies have been limited to less-stressful scenarios by human subject protocols and privacy considerations. In these studies, subjects are motivated to deceive primarily by financial reward.

Our collection paradigm was designed to elicit within each subject deceptive and nondeceptive speech, from subjects who had both financial incentive and motivation in terms of what De-Paulo [2] calls the 'self-presentational' perspective to do well at deception. Thirty-two native speakers of Standard American English were recruited for the study. Subjects were asked to perform a series of tasks (activities and question answering) in six areas, and were told that their performance would be compared to a target profile based on a survey of the twenty-five 'top

entrepreneurs of America' performing similar tasks, results of which they would be shown later. Task difficulty was manipulated so that subjects scored more poorly than the target in two task areas, better than the target in two others, and the same in another two of the six; this manipulation was balanced across task categories.

In the next phase of the experiment, subjects were shown their own score and the target, which were invariably quite different in four areas. They were told that the study's actual goal was to compare people who have certain skills and knowledge with people who are good at convincing others that they do. They were told that they could continue to the second stage of the study and also be eligible for a $100 prize if they could convince an interviewer that, instead of scoring as they had, they had in fact performed just as the target entrepreneurial profile.

Thus, each subject was motivated to tell the truth in two task areas and to deceive the interviewer in four others. They were told that the interviewer had no knowledge either of the target profile or of their performance (the latter true). The interviewer's task was to determine how he thought the subjects had actually performed, and he was allowed to ask them any questions other than those that were actually part of the tasks they had performed. Finally, for each question, subjects were asked to indicate whether the reply was factually true or contained any false information by pressing one of two pedals hidden from the interviewer under the table. Although we are unaware of studies addressing the impact of pedal pressing on speech production, we attempted to address this issue in two ways. First, subjects pressed a pedal in both truth and lie conditions, so any impact should affect both conditions equally. In addition, subjects were asked if they found it difficult to use the pedals, and the vast majority responded no.

The interviews, which lasted between 25 and 50 minutes, comprised 15.2 hours of interviewer/subject dialog and yielded approximately 7 hours of subject speech. They were recorded to digital audio tape on two channels using a Crown CM311A Differoid headworn close-talking microphone and downsampled to 16 kHz. They were subsequently orthographically transcribed, and sentence-like units ("slash units", or SUs) [3]) were labeled. The transcription was then automatically aligned with the audio data. The SUs are one candidate unit chosen from many others (e.g. units defined between pauses, turn based units, etc). Further experiments could be done to compare classification results using different units.

3. FEATURES AND CLASSIFIERS

Previous research and practitioner experience suggest that acoustic-prosodic and lexico-syntactic cues may signal when speakers are deceptive. Below we describe the lexical and acoustic-prosodic cues we used in our corpus.

3.1. Prosodic-Lexical SVM System

Observations in the literature suggest that pitch, energy, speaking rate, and other stylistic factors (e.g., "muffled" voice) vary when speakers deceive. Our prosodic features attempt to capture this variation as well as to explore other potential cues. We considered a wide range of potential acoustic and prosodic features, taking advantage of tools available from automatic speech recognition, to extract and model features including duration, pausing, intonation, and loudness, associated with multiple time scales, from a few milliseconds to an entire speaker turn. Prosodic features are automatically normalized, taking into account long-term speaker-specific habits as well as segmental context.

To extract **prosodic features**, the speech was first segmented into SUs by chopping at punctuation marks (ellipses, periods, and question marks) in the hand-transcribed corpus. For each SU, we computed 215 prosodic features involving pitch, energy, and duration patterns. Pitch and energy were obtained from the ESPS/Waves pitch tracker *get_f0*; duration features were obtained via forced alignment of hand transcripts using the SRI automatic speech recognition system. Pitch features were computed from the voiced regions in the SU, and were then used in one of three forms: raw, median-filtered, or stylized using an approach that fits linear splines to the median-filtered pitch. From these pitch sequences we computed a large set of features, including maximum pitch, mean pitch, minimum pitch, range of pitch number of frames that are rising/falling/doubled/halved/voiced, length of the first/last slope, number of changes from fall to rise, and value of first/last/average slope. These features were normalized by five different approaches: no normalization, division by the mean, subtraction of the mean, and *z-scores* (subtracting the mean and dividing by the standard deviation). Two basic energy features were computed. The first was the raw energy in the SU and the second was the raw energy of only the voiced regions. The second feature type was used in one of three forms: raw, median-filtered, or stylized using a linear spline-fitting approach. From these values we computed several derived features, including the maximum energy, minimum energy, mean energy, and other features similar to those just mentioned for pitch. Finally, several duration features were computed. The maximum and the average phone duration in the SU were first computed. They were then used either as raw values, normalized using speaker-specific durations or normalized using durations computed from the whole corpus. The corpus-based normalization was done dividing by the mean, or subtracting the mean and dividing by the standard variation.

Lexical features were computed automatically using true words from hand transcriptions. These features were based on results or hypotheses from the literature [2] and on intuitions of practitioners in the intelligence and law enforcement communities. They include counts of filled pauses, syntax-based features, dialog act labels such as specific denials, flags for positive and negative emotion words [13], and a feature encoding whether a subject responded to the interviewer's question with a question. For each SU, we computed 20 lexical features. This is a preliminary set of features and we believe further gains can be achieved by adding more lexical features.

A support vector machine (SVM) classifier with a linear kernel was used with the prosodic-lexical features. The total input feature dimension was 235 for the prosodic/lexical SVM system and 215 for the prosodic SVM system. We used the freely available LIBSVM tool [12] for training and testing the SVM. A zero mean and unit standard deviation normalization was used with the input features. Other kernels (radial basis and polynomial) were also tried, but we found the linear kernel to give the best results.

One problem for SVMs is the missing features for some SUs (for example, the maximum positive slope feature in a short unit with a negative slope). We found very few cases of missing features in our CSC corpus. Missing feature values were replaced by the mean of the observed values for that feature.

3.2. Acoustic GMM System

The acoustic system was built to discriminate truthful and deceptive speech using features computed in the spectral domain. This model is similar to the one used in speaker identification systems [9].

The features used in this system are spectral-based Mel cepstral features with energy, plus simple, double and triple delta features. The total feature dimension is 52. The acoustic features were computed from 25 ms Hamming-windowed signal frames, stepped every 10 ms. The signal energy (C0) was normalized by the maximum over the complete waveform. In order to avoid using silent or noisy frames, we used only frames whose energy was at least a minimum difference over the maximum signal energy.

A Gaussian mixture model (GMM) classifier was used with the acoustic features. The total number of Gaussians in the Gaussian mixture was 2048. First, a boot GMM was trained using the expectation maximization (EM) algorithm to maximize the likelihood on the training data. This boot model was trained with all the training data from both classes (truthful and deceptive). Next, two different GMMs were created by adapting the boot GMM to the truthful section and to the deceptive section of the training data. The adaptation algorithm was the maximum *a posteriori* adaptation (MAP) algorithm. A class decision is produced by this system comparing the class posterior probabilities from each GMM for a given waveform (using priors estimated from the training data). By adapting both target models from the same boot model we ensure that the likelihood scores are comparable.

3.3. Combiner SVM System

The purpose of the combiner was to evaluate whether combining scores from both systems would improve the classification accuracy. The rationale is that if each system provides a confidence measure for its class prediction, the combiner will weight the evidence from each system and thus may improve the class prediction. The score combiner was an SVM with a radial basis kernel.

The score generated from the acoustic GMM system was the ratio of the truthful GMM posterior probability and the deceptive GMM posterior probability. The class priors were estimated from the training data. The score generated from the prosodic-lexical SVM system was the output of the dot product between the kernel output of the support vectors and the kernel output of the input vector. This corresponds to the signed distance in kernel space of the test data point from the decision boundary. We also tested a 3-way combination of the GMM, the prosodic SVM and the lexical systems but did not find an improvement over the 2-way combination.

The combiner was trained on a subset of the training data. We split the training data into two sets we will call devtrain and devtest. The split proportion was 80% for training and 20% for testing. The prosodic-lexical SVM and the acoustic GMM were retrained in the devtrain data. The scores from each system were generated for the devtest data, and the combiner was then trained on that data. For independent testing the two systems were retrained on the full training set. A zero mean and unit standard deviation normalization was used with the scores from each system. The normalization parameters were computed from the devtrain data and were applied to the test data.

4. EXPERIMENTS

We first explored the performance of each system, and then the performance of the combined system. We finally assessed the effects of word recognition errors by evaluating the prosodic system using features computed from the recognized words instead of transcriptions.

4.1. Data

Each speaker's SUs were partitioned into 90% for training and 10% for testing. Then the training data from all speakers was pooled to form the final training data. The same procedure was used for the test data resulting in a total of 8406 training SUs and 922 test SUs. Before splitting the data by speaker, a randomization of the SUs was done with the same seed for all speakers. The collection of training data and test data randomized with the same seed was called a "run". Ten different runs were produced, each with a different seed. All results represent averages over the 10 different runs. The pedal press information was used to assign a truth or lie label to each SU.

Since the same speakers occur in both training and test, our experiments are speaker dependent, thereby allowing the expected speaker-dependent effects to be modeled. However, we decided to pool data from all speakers since the amount of data per speaker experiment would otherwise be insufficient for effective model training.

4.2. Results

Table 1 presents accuracy results of the acoustic GMM system, the prosodic/lexical SVM system and the prosodic only SVM system in the CSC Corpus test data partition, as well as results for the combined systems. The chance result is simply the ratio of the more frequent class (truth) to the total number of units in the test set, corresponding to a classifier that ignores the test data and always outputs the *a priori* most likely class.

Table 1: Accuracy of Single Systems and Combination Systems on the CSC Corpus.

Systems	% Accuracy
Chance	60.4
(A) Acoustic GMM	62.1
(B) Prosodic SVM	62.7
(C) Prosodic/Lexical SVM	62.9
Combination of Systems A and B	64.4
Combination of Systems A and C	64.0

From Table 1 we conclude that each individual system produces a gain over chance, and the prosodic-based systems produce the largest gains. Adding the lexical features to the prosodic features gives higher accuracy than the prosodic features alone. The combination of systems A and B produces the best accuracy and the combination of systems A and C results in a similar performance. One reason why the combination of systems

A and C was not better than the combination of systems A and B may be that by adding the lexical features, both systems become more similar, with fewer different errors for the combiner to leverage. The difference between these two combiners is not statistically significant.

When a matched pairs test is used the difference in accuracy between chance and the combination of systems A and B is significant ($p < 0.05$) and the difference in accuracy between chance and the combination of systems A and C is also significant ($p < 0.10$).

The combiner is an SVM with a radial basis kernel. We explored using other kernels such as linear and polynomial. The linear kernel produced no gain over the individual systems. The third-degree polynomial kernel when used in combination of systems A and B produced an intermediate accuracy of 63.9%. Thus, the radial basis kernel outperforms the polynomial kernel, which in turn outperforms the linear kernel, showing that the class boundaries in the combiner are nonlinear in shape.

4.3. Prosodic System from Recognized Words

Finally, we compare accuracy results from the prosodic-only SVM system using features computed from automatically recognized words versus features computed from human transcriptions. This is important for feature extraction on untranscribed input, where only recognized words are available.

The same original SU boundaries from the previous experiment were used. We used a conversational telephone speech recognizer adapted for full-bandwidth recordings [10]. The same procedure of data splitting and 10 run repetition was used as before. Since some short utterances could not be recognized, the total number of SUs was 874 for testing and 8104 for training. In Table 2 we present the accuracy of both systems on the CSC corpus.

Table 2: Accuracy of Prosodic SVM Systems using Features from Transcripts and Recognizer in Smaller Train and Test Sets on the CSC Corpus.

Systems	% Accuracy
Chance	60.4
Prosodic SVM from Recognized Words	62.6
Prosodic SVM from Transcripts	62.8

From Table 2 we conclude that the prosodic SVM system using features extracted from recognized words performs similarly to the same system but using features extracted from the true words; the difference is not statistically significant.

The comparison in Table 2 reveals a reasonable lack of sensitivity of the prosodic features to recognition errors. The most sensitive features were probably phone-based features (i.e. phone durations) and rate-of-speech based features (i.e. number of words divided by the SU duration).

5. CONCLUSIONS

In this paper we have described experiments on distinguishing deceptive from non-deceptive speech in the CSC Corpus. Specifically we have proposed a system combination approach which provides greater accuracy than the individual systems. The experimental results reveal that there is potential for further improvement by adding more independent systems. Additionally we began to explore the impact on accuracy of features computed from recognized words. Future work will focus on improving the individual systems by adding voice quality features and exploring other acoustic front-ends, on developing new features that are less sensitive to recognition errors, and on proposing new independent systems.

6. ACKNOWLEDGEMENTS

This research was supported by grants from the National Science Foundation (NSF IIS-0325399) and the Department of Homeland Security.

7. REFERENCES

[1] S. H. Adams, "Statement Analysis: What Do Suspects' Words Really Reveal?," *FBI Law Enforcement Bulletin*, October 1996.

[2] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to Deception," *Psychological Bulletin*, 129(1):74–118, 2003.

[3] NIST. Fall 2004 Rich Transcription (RT-04f) Evaluation Plan, August 2004. http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.pdf.

[4] P. Ekman, M. Sullivan, W. Friesen, and K. Scherer, "Face, Voice, and Body in Detecting Deception," *Journal of Nonverbal Behaviour*, 15(2):125–135, 1991.

[5] D. Haddad and R. Ratley, "Investigation and Evaluation of Voice Stress Analysis Technology," March 2002. National Criminal Justice Reference Service, http://www.ncjrs.org/pdffiles1/nij/193832.pdf.

[6] J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, E. Shriberg, and A. Stolcke, "Distinguishing Deceptive from Non-Deceptive Speech," in *Proc. Eurospeech 2005*, Portugal, pp. 1833–1836, September 2005.

[7] A. Mehrabian. Nonverbal Betrayal of Feeling. *J. Experimental Research in Personality*, 5:64–73, 1971.

[8] J. E. Reid and Associates. *The Reid Technique of Interviewing and Interrogation*. Reid, John E. and Associates, Inc., Chicago, 2000.

[9] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Mixture Models", *Digital Signal Processing*, vol. 10, pp.181-202 (2000).

[10] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng (2005), "Further Progress in Meeting Recognition: The ICSI-SRI Spring 2005 Speech-to-Text Evaluation System, *Proc. NIST MLMI Meeting Recognition Workshop*, Edinburgh.

[11] L. A. Streeter, R. M. Krauss, V. Geller, C. Olson, and W. Apple, "Pitch Changes during Attempted Deception," *Journal of Personality and Social Psychology*, 35(5):345–350, 1977.

[12] C.-C. Chang and C.-Jen Lin, LIBSVM: A Library for Support Vector Machines, 2001. www.csie.ntu.edu.tw/~cjlin/libsvm

[13] C. Whissell. The Dictionary of Affect in Language. In R. Plutchnik and H. Kellerman, editors, *Emotion: Theory, Research and Experience*, pages 113–131. Academic Press, New York, 1989.