A STUDY OF GLOTTAL WAVEFORM FEATURES FOR DECEPTIVE SPEECH CLASSIFICATION

Juan F. Torres, Elliot Moore II, Ernest Bryant

Georgia Institute of Technology
School of Electrical and Computer Engineering
210 Technology Circle, Savannah, GA, 31407

ABSTRACT

Previous work in detection of deceptive speech has largely focused on prosodic, vocal tract, and lexical features. Glottal waveform features have been shown to be useful discriminators for various types of speaker affect and warrant further study within the context of deception detection. This paper reports on speaker-dependent machine learning and feature selection experiments for classifying deceptive and non-deceptive speech using a large number of statistical features derived from the glottal waveform. We present current results comparing the classification performance and selected feature sets across 19 speakers from the Columbia-SRI-Colorado corpus of deceptive speech and discuss directions for future work.

Index Terms— Speech Analysis, Feature Extraction

1. INTRODUCTION

The growing focus on security as a worldwide problem has stimulated the development of technology that can detect potential threats in an objective and non-intrusive manner. One area of interest is the ability to automatically detect willful deception, with the speech modality being particularly valuable since it is ubiquitous and easily collected. Several Voice Stress Analyzers (VSA) have been implemented by law enforcement [1] based on a theory known as "vocal tremor"; however, no single tool has been developed that is capable of directly identifying deception. Instead, the detection of willful deception is based on the interpretation of an increase in perceived stress in a subject. In light of this, the work presented here reflects an effort to carefully identify ways to measure this stress in speech and correlate it to known instances of willful deception.

Several recent studies using the Columbia-SRI-Colorado (CSC) corpus of deceptive speech have explored the possibility of automatic deception classification using lexical, prosodic, and MFCC features: Gracierena et al. [2] trained a classifier with a combination of MFCC and prosodic features from a pool of 32 speakers, obtaining 64.4% accuracy out of a baseline (chance) score of 60.4%, while Hirschberg et al. [3] obtained 62.8% accuracy out of a 60.2% baseline using a combination of lexical and prosodic features, with a further improvement of 3.6% when speaker-dependent features and gender were used. Enos et al. [4] obtained a much improved accuracy of 61.9% out of a 50% baseline when classifying global lies (Section 2) in the corpus using "critical segments" of speech. The results in [3] are consistent with the observation by Ekman et al. [5] that there may be significant differences in the cues elicited by individual subjects during deception.

To our knowledge, no previous studies have attempted to classify deceptive speech using features derived from the glottal waveform. However, glottal waveform features have been shown to be useful discriminators for other types of speaker affect, including stress and styling [6, 7], simulated emotion [8], and depression [9]. These results suggest that there may be significant differences in vocal fold vibration during deceptive and non-deceptive speech that cannot be completely characterized by pitch information and motivate the study of glottal waveform features for deception classification. This paper reports on speaker-dependent machine learning and feature selection experiments for classifying deceptive and non-deceptive speech from the CSC corpus using a large number of statistical features derived from the glottal waveform. The paper is organized as follows: In Section 2 the CSC corpus is briefly described. The glottal waveform estimation and feature extraction procedure is described in Section 3. Classification and feature selection experiments are described in Section 4. Results are discussed in Section 5 and conclusions are presented in Section 6.

2. THE CSC DECEPTION CORPUS

The Columbia-SRI-Colorado (CSC) corpus [3] contains deceptive and non-deceptive speech from 32 native speakers of

This work was supported in part by the National Science Foundation (Grant No. 0545772). The authors kindly thank Julia Hirschberg, Stefan Benus, Jason M. Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, Bryan Pellom, and Elizabeth Shriberg for access to the CSC Deceptive Speech Corpus.

Standard American English in an interview setting. Interviewees were given a set of tests and later told that their performance in certain test sections did not match the target profile, but that the study also sought subjects who could deceive the interviewer. They were told that successful deceivers would qualify for an additional $100. Thus, the interviewees were under a financial and 'self-presentational' incentive to deceive the interviewer. The veracity of each statement made during the interview was indicated by pressing one of two pedals hidden beneath the table. This interview procedure resulted in the production of two kinds of lies: Global lies describe the speaker's overall intention to deceive with respect to their performance in a particular test section, while local lies refer to individual deceitful statements made to support the overall argument. The interviews lasted between 25 and 50 minutes, and the corpus contains about 7 hours of subject speech. The interviews were recorded into digital audio tape using a high-quality head worn microphone and downsampled to 16 kHz. The contents of the corpus are fully described in [3].

3. GLOTTAL FEATURES

The glottal waveform is defined as the volume velocity of airflow at the back end of the vocal tract. An estimate of the glottal waveform may be obtained from the speech signal via inverse-filtering techniques, which attempt to estimate and remove the effects of vocal tract resonances. From these estimates, 13 time and frequency-domain glottal parameters were computed at the frame level. Each parameter was then subjected to a set of 10 statistical measures across the frames of each continuous voiced section, producing a large set of statistical features that were used for classification.

Feature extraction proceeded as follows: First, the VOICEBOX [10] implementation of the RAPT pitch estimation algorithm was used to segment the speech in the CSC corpus into continuous voiced sections. The vo_bias parameter was manually adjusted between -0.4 and -0.6 on an individual speaker basis so that only strongly-voiced speech frames were selected, as verified by auditioning the speech tagged as voiced and unvoiced. The length of the extracted voiced sections varied between 70 ms and approximately 2 seconds, with a mean length of 255 ms.

Glottal closure instants (GCI's) were obtained using the DYPSA algorithm [11]. Voiced sections were divided into 50%-ovarlapped frames of length equal to 3.5 times the mean pitch period across the entire voiced section, but restricted to values between 20 and 40 ms. Glottal waveform estimates were obtained for each frame using the Rank-Based Glottal Quality Assessment (RBGQA) algorithm [12], which combines four glottal estimate quality measures to find the optimal analysis window position for deconvolution via the covariance method of linear predictive analysis (LPA). An LPA order of 14 was used throughout.

To explore a wide range of potential deception cues in the glottal domain, we computed (for each frame) the 13 glottal waveform parameters implemented in version 0.3.0 of the APARAT toolbox [13], which have been shown in the literature to be related to various aspects of voice quality. These parameters are listed in Table 1; further references may be found in the APARAT documentation. The following statistical measures were computed on each glottal waveform parameter vector, as well as on rectified and unrectified versions of their delta and delta-delta vectors: mean, median, min, max, standard deviation, range, dynamic range, interquartile range, skewness, and kurtosis. This procedure resulted in 23643 observations (9376 lie, 14267 truth) at the voiced-section level, each containing 650 feature statistics. Each observation was labeled using lie/truth labels at the local lie level (Section 2). The mean number of observations per speaker was 739 (293 lie, 446 truth).

Table 1. Glottal Waveform Parameters

aq	Amplitude quotient
clq	Closing quotient
H1-H2	Difference between 1^{st} and 2^{nd} glottal formants, in dB
hrf	Harmonic richness factor
naq	Normalized amplitude quotient
oq1	Open quotient, calculated from the primary glottal opening
oq2	Open quotient, calculated from the secondary glottal opening
oqa	Open quotient, derived from the LF model
pch	Pitch, calculated from the distance between GCI's
psp	Parabolic spectrum parameter
qoq	Quasi-open quotient
sq1	Speed quotient, calculated from the primary glottal opening
sq2	Speed quotient, calculated from the secondary glottal opening

4. EXPERIMENTS

In order to study speaker-dependent glottal effects in deceptive speech, all experiments were performed separately on each speaker. An initial reduction in feature set size was obtained using a Kolmogorov-Smirnov (K-S) test to estimate significant variations in single features with respect to class. Features whose distributions did not differ to a significance level of $p < 0.2$ were discarded.

After the removal of irrelevant features, sequential forward floating selection (SFFS) [14] was performed in a wrapper approach. SFFS starts with an empty feature set and adds a single optimal feature in each forward iteration until classification accuracy no longer improves. At this point, the algorithm attempts to remove a single feature in each backward

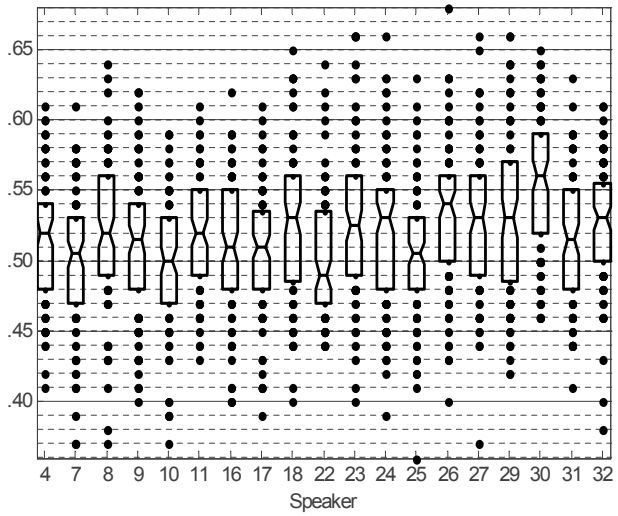

Fig. 1. Box plots showing the distribution of classification accuracy for each speaker across 100 runs. Horizontal lines indicate medians; notches denote comparison intervals for the medians at a 95% significance level; boxes show interquartile (25^{th} – 75^{th} percentile) ranges; Results beyond the middle 50% are shown as black dots.

iteration while classification improves. The algorithm stops when the forward and backward stages fail to improve classification accuracy. Within each SFFS iteration, a Gaussian Mixture Model (GMM) classifier was trained via the Figueiredo-Jain (FJ) algorithm [15], which automatically chooses the optimum number of mixtures during training.

For each speaker, a random subset of 200 observations per class (truth/lie) was selected and divided into 100 observations (per class) for GMM training, 50 for testing inside the SFFS iterations, and 50 for validation. The validation set was used to obtain a final score after SFFS had selected a final feature set. This configuration allowed for a fair performance comparison between speakers, since the baseline score was always 50% and each speaker had the same amount of data for model training. The independent validation set was necessary to obtain an unbiased score, since the testing set is already used within the feature selection procedure [14].

Each randomly-selected subset of 200 observations (with a different seed) was called a "run," and 100 SFFS runs were performed for each of 19 speakers. The remaining 13 speakers in the CSC corpus were not analyzed at this time because they contained less than 200 observations in either class.

5. RESULTS

The classification accuracy distributions for each speaker, given as box plots in Figure 1, show wide variations across runs. There are two main reasons for this: Firstly, like all sequential feature selection procedures, SFFS is susceptible to convergence on local minima, which suggests that the better runs may be showing the true classification potential of the glottal features. However, it must also be noted that although the validation dataset is completely disjoint from the training and test sets, the random sampling procedure for each run may result in data partitions in which the test and validation sets contain many similar (or dissimilar) observations, which may further influence the final validation score.

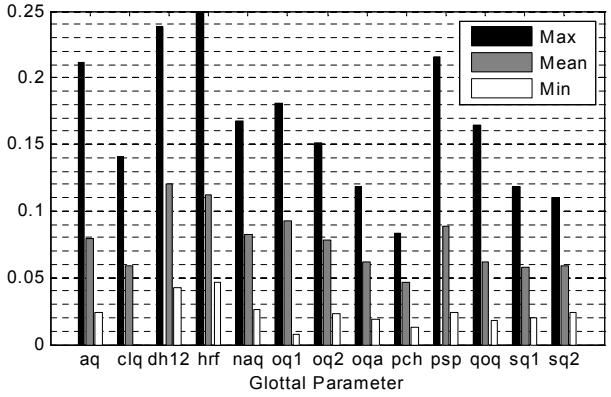

Fig. 2. Normalized selection frequency of glottal parameters across speakers, computed from runs at or above the 75^{th} percentile. The figure shows the maximum, mean, and minimum values across speakers.

Taking these issues into consideration, conclusions about classification performance were derived from the statistical properties of the distributions, rather than from the best score in an individual run: From Figure 1, it can be seen that 17 out of the 19 speakers obtained median scores slightly above chance. While the results for most speakers are too close to chance to draw definite conclusions, speakers 26, 30, and 32 scored above chance in approximately 75% of the runs (25th percentile), strongly suggesting the presence of glottal effects associated with the deceptive speech of these speakers.

As a preliminary investigation into the discriminatory ability of each glottal parameter, we produced selection frequency histograms for the 13 glottal parameters listed in Table 1, using the final feature sets from runs having a score at or above the 75^{th} percentile. Assuming that the feature selection procedure is more likely to select useful features than irrelevant features, the selection frequency of feature statistics derived from a particular glottal parameter can be interpreted as a rough indication of the parameter's relative usefulness.

Given the 13 glottal parameters, a random probability of selection of $p \approx 0.077$ is assumed. A summary of selection frequency across all speakers is given in Figure 2. What mainly stands out from this figure is the consistently low selection frequency of the pitch parameter, which supports the hypothesis that vocal affect related to deception may be more clearly manifested as variations in glottal waveform shapes and spectra than as a simple change in pitch.

Examination of the individual feature selection histograms for each speaker (not shown) revealed $H1 - H2$ to be the most consistently-selected parameter across speakers, with a selection frequency above $1.5p$ for 10 out of the 19 speakers. While other glottal parameters were selected with high frequency for certain small groups of speakers, feature selection patterns varied widely across speakers, as evidenced by the large distances between the maximum, mean, and minimum values in Figure 2.

6. CONCLUSION

This paper has described experiments for classifying deceptive and non-deceptive speech on an individual speaker basis using features derived from the glottal waveform. The results presented here highlight a few critical points about deception in speech: One is that the expression of deception in the glottal domain is highly varied across speakers, which makes the determination of any global feature set problematic. Additionally, while several runs demonstrated performance "above chance," the overall median performance of the classifiers must be greatly improved for practical use. It should be emphasized, however, that these results are from experiments that attempted to classify deception at the voiced-section level, which typically consisted of 100–300 ms of speech. While this observation level was useful for studying the speaker-dependence of glottal effects associated with deception, more useful classification results may be obtained by classifying entire local lie statements or global lie 'critical segments' [4]. Finally, given that this feature set is limited to a single domain of speech analysis, there is additional encouragement that the combination of other acoustical and lexical domains can improve overall classification accuracy.

Future work includes the analysis of a larger set of glottal parameters, the development of classification experiments at longer observation levels, and the combination of glottal features with other types of acoustic and lexical features.

7. REFERENCES

[1] F. Horvath, "Detecting deception: the promise and the reality of voice stress analysis," *Journal of Forensic Sciences*, vol. 27, no. 2, pp. 340–351, 1982.

[2] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *IEEE Int. Conf. Acous. Spch. Sig. Process.*, 2006, vol. 1.

[3] 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 *Interspeech*, 2005.

[4] F. Enos, E. Shriberg, M. Graciarena, J. Hirschberg, and A. Stolke, "Detecting deception using critical segments," in *Interspeech*, 2007.

[5] P. Ekman, M. Sullivan, W. Friesen, and K. Scherer, "Face, voice, and body in detecting deceit," *Journal of nonverbal behavior*, vol. 15, no. 2, pp. 125–135, 1991.

[6] K. Cummings and M. Clements, "Analysis of the glottal excitation of emotionally stressed speech," *J. Acoust. Soc. Am.*, vol. 98, pp. 88–98, 1995.

[7] J. H. L. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," in *Speaker Classification I*, C. Muller, Ed., 2007, vol. 4343 of *Lecture Notes in Artificial Intelligence*, pp. 108–137.

[8] M. Airas and P. Alku, "Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient," *Phonetica*, vol. 63, no. 1, pp. 26–46, March 2006.

[9] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Transactions on Biomedical Engineering : Accepted for future publication*, 2007.

[10] M. Brookes, "Voicebox: Speech processing toolbox for matlab," 2007, `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

[11] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Speech Audio Lang. Processing*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[12] E. Moore and J. Torres, "A performance assessment of objective measures for evaluating the quality of glottal waveform estimates," *Speech Communication (in press)*, 2007.

[13] M. Airas, H. Pulakka, T. Backstrom, and P. Alku, "A toolkit for voice inverse filtering and parametrisation," in *INTERSPEECH*, 2005, pp. 2145–2148.

[14] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature Extraction: Foundations and Applications*, Springer, The Netherlands, 2006.

[15] M. Figueiredo and A. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 3, pp. 381–396, 2002.