INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Acoustic Indicators of Deception in Mandarin Daily Conversations Recorded from an Interactive Game

Chih-Hsiang Huang, Huang-Cheng Chou, Yi-Tong Wu, Chi-Chun Lee, Yi-Wen Liu

Department of Electrical Engineering, National Tsing Hua University, Taiwan

carryjim@gapp.nthu.edu.tw, hc.chou@gapp.nthu.edu.tw, tonetone8513@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw, ywliu@ee.nthu.edu.tw

Abstract

Being able to distinguish the differences between deceptive and truthful statements in a dialogue is an important skill in daily life. Extensive studies on the acoustic features of deceptive English speech have been reported, but such research in Mandarin is relatively scarce. We constructed a Mandarin deception database of daily dialogues from native speakers in Taiwan. College students were recruited to participate in a game in which they were encouraged to lie and convince their opponents of experiences that they did not have. After data collection, acoustic-prosodic features were extracted. The statistics of these features were calculated so that the differences between truthful and deceptive sentences, both as they were intended and perceived, can be compared. Results indicate that different people tend to use different acoustic features when telling a lie; the participants could be put into 10 categories in a dendrogram, with an exception of 31 people from whom no acoustic indicators for deception were found. Without considering interpersonal differences, our best classifier reached an F1 score of 53.37% in distinguishing deceptive and truthful segmentation units. We hope to present this new database as a corpus for future studies on deception in Mandarin conversations.

Index Terms: deception, trust, computational paralinguistics, Mandarin, corpus

1. Introduction

Deception analysis and detection not only is a significant skill when dealing with financial fraud related scenarios [1], but also plays an irreplaceable role in our daily dialogues with other people. Among features of different modalities, the acoustic-prosodic features have been used for deception analyses in many works. Almost all of the existing databases are in English [2, 3], probably because up to now deception-related research is predominantly driven by Western countries. It was also argued that more attentions needed to be put on the automatic deception detection in Asian languages [4]. The cultural differences and ethical considerations [4] tightly correlate the behaviour of deception, therefore we hope to discover how acoustic cues are demonstrated from the interaction between Mandarin native speakers.

A Mandarin database was constructed by collecting deceptive speech from a game [5, 6]. Participants were asked to prepare either a fake or a truthful story about themselves and present it in front of two interviewers. In this work, we intended to collect deceptive speech from open conversation speech and the resulted data are referred to as daily deceptive dialogues (DDD).

Daily deceptive dialogues can be encountered everywhere in our life but little research focused on it because the data are hard to collect and tricky to handle at the labeling stage. In this work, an interactive game was designed for DDD data collection and we were able to collect about 27.2 hours of the recording data. The subjects are gender-balanced and are all native speakers of Mandarin with different degrees of Taiwanese accents.

During the experiment, the subjects were assigned to lie about one or two out of three topics according to a questionnaire they filled when being recruited. Then, each of the game sessions would invite two subjects to chat about all the topics, and their job was to deceive their opponents. We are interested in what combinations of acoustic features that the subjects might have utilized, consciously or not, to convince their opponents about experiences that they did not have.

The remaining sections are organized as follows. Related research is reviewed in Sec. 2. The detailed procedure of construction of our Mandarin DDD corpus is described in Sec. 3. Sec. 4 statistically analyzes the corpus based on acoustic features and Sec. 5 describes our preliminary results of the automatic deception detection. Discussion and conclusions are given in Sec. 6.

2. Related work

This topic has been discussed for several years and a persisting dilemma is that, at the stage of data collection, the subjects might be influenced by fear and stress and thus lack the spirit of lying [7]. Therefore, the existing English deceptive corpora paid much effort at the stage of collecting data [2, 8]. The Interspeech 2016 ComParE Deception Sub Challenge [3] provided Deceptive Speech Database (DSD) and a baseline acoustic feature set [9, 10] achieved an unweighted average recall (UAR) score of 68.3%. This baseline feature set consists of statistics from various functionals over low-level descriptor (LLD) contours [11, 12]. Levitan et al. [13] combined the baseline feature set with LIWC [14], DAL [15] and phonotactic variation features to reach 69.4% on the DSD testing set through the sequential minimal optimization algorithm. In addition, they trained and evaluated across corpus with Columbia Deception Corpus (CDC) [2] and suggested that the acoustic-prosodic features do generalize and are promising for deception detection. Sondhi et al. [16] also showed that under stressful circumstances when telling lies, the emotional and cognitive states experienced by the subjects would influence their acoustic features.

For the spoken language Mandarin, Cheng Fan et al. constructed the SUSP-DSD corpus [5, 6] through a three-phase experiment designed to obtain deceptive descriptions in different scenarios; however they only used the third phase for evaluation due to the difficulties about the setting of baseline. Another limitation of their work is that the scale of the SUSP-DSD corpus is smaller when comparing with the existing English corpora. We know that the behaviours of deception diverge among individu-

http://dx.doi.org/10.21437/Interspeech.2019-2216

Figure 1: *The recording environment. Our member (middle) was instructing the subjects about the game. During the game, our member would leave and monitor the room from outside.*

als. Therefore, the scale of the deception corpus is an indispensable element especially for the purpose of generalization. Xiaohe Fan et al. [17] further extracted Mel-frequency cepstrum coefficient (MFCC) and zero crossing rate (ZCR) from SUSP-DSD corpus as features for K-SVD algorithm. This sparse coefficients based algorithm reached an accuracy of 72.95% which is better than their previous work.

3. DDD corpus of Mandarin

3.1. Subject recruiting

In order to minimize emotional effects from the environment and the experiment itself, we designed the experiment as a human-interactive game. All of the subjects are native speakers of Mandarin with Taiwanese accent and are aged from 20 to 25. When recruiting, the subjects were picked based on whether they had sufficient experiences about the questions on the questionnaire, which was the same questionnaire they would use when playing the game. The way each question on the questionnaire were carefully crafted so as to minimize the chance of eliciting ambiguous answers during the game such as, 'Oh, I forgot when I attend the club' or 'It has been a long time so I forgot the details'. The following are our three questions: 'Have you ever competed in ball games?', 'Have you ever attended any music instrument competitions, or performed in a concert after senior high school?'. 'Have you ever attend any performing club and its final presentation?'. In total, 100 subjects were hired, consisting of 50 male and 50 female subjects. The design of the game was approved by the IRB of National Tsing Hua University (Record No. 10612HE092).

3.2. Daily deceptive dialogue corpus of Mandarin

Each session of the game was participated by two subjects who had never met before the game. The questionnaire contained 3 questions. Both subjects were assigned one or two deceptive questions according to the questionnaire they filled in when recruiting. Here, we assumed that deception should not be too difficult because DDD appears in daily life, and this is opposite to the designing spirit of [2]. To encourage the subjects to convince their opponent, we provided a secret gift to the side who earned the most trusts on these three questions. Data from 4 out of 100 subjects had to be abandoned due to recording problems. In the end, we collected about 27.2 hours of clean recording data consisting of 14 pairs of female-female, 14 pairs of male-male and 20 pairs of female-male subjects. The following analyses are based on these 96 subjects.

3.3. Recording environment and equipment

The experiment was conducted in an sound-proof and sound absorbing studio (Fig.1). In each session, two subjects were instructed to sit on the chair face to face without separated by a curtain. Two directional microphones (SHURE SM58) were pointed to the subjects individually and were fixed by stands. The recording interface was MOTU UltraLite-mk4 sound card with Cubase Pro 9.0 recording software. The format audio was 24-bit, mono channel with a sampling frequency of 44.1 kHz.

3.4. Segmentation unit organization and labeling

The segmentation unit in this paper is sentence-like unit (SU) which is different from the inter-pausal unit (IPU) of [2]. We think IPU is not suitable for our corpus because if the subjects stammer or pause, this kind of method loses information and is inconsistent. As a result, if we define the side who raise questions as *interviewer* and the side who responses as *interviewee*, the responses of each subject form a set of SUs, which have been manually organized by the following indices: $i = 1, 2, 3$ is the index for questions on the questionnaire, and $j = 1, 2, ..., m(i)$ is the index for subquestions raised by the interviewer for the ith question on the questionnaire. Let $n = n(i, j)$ denote the number of SUs from one interviewee to each subquestion. Then, for each i and j, we have organized the set of corresponding SUs as follows,

$$q_{ij} = \{SU_k\}_{k=1}^{n(i,j)}, \tag{1}$$

where SU_k denotes the kth SU in q_{ij}. Then, the set of all SUs responded by a subject to question $i = 1, 2$ or 3 can be defined,

$$Q_i = \{q_{ij}\}_{j=1}^{m(i)}. \tag{2}$$

Finally, the set of all the responses from a subject can be expressed as $r = \{Q_1, Q_2, Q_3\}$. Thus, all the SUs are stored in different files under different folders.

Two kinds of labeling were proposed in [2, 18], one is local deception and the other is global deception. It is known that not all of the components are deceptive when a lie is told. Therefore, local deception defines the ground truth of each response while global deception counts all responses toward one question as the same label. Due to the design of the present experiment, we could only adopt global deception as our labeling method. Nevertheless, in addition to the ground truth of each question provided by the interviewees, we also collected the interviewers' guessed answers. In this paper, the number of the truth/deception SUs in all, male and female subjects are 2356/2082, 1087/1018 and 1269/1064, respectively; the number of perceived truth/deception of SUs in all, male and female subjects are 2914/1524, 1574/542 and 1340/982, respectively.

3.5. Eysenck Personality Questionnaire (EPQ)

Participants were all asked to fill an EPQ after they finished the experiment. EPQ has been slightly modified according to the cultures and customs of different countries [19]. The EPQ we used in this experiment contains 85 items which was edited from [20]. EPQ is used to assess the personality traits of a person from the following four factors. Extraversion (E): Those who score high in this dimension show the characteristics of outgoing, talkative and desire to explore. Those who score low tend to have a stable emotion, stay distant to people except intimate individuals and lead a regular life. Neuroticism (N): It is characterized as a normal behaviour instead of symptoms. High

Table 1: *A list of common acoustic indicators for each group clustered in Fig. 2, respectively. Group A to J represent the 10 branches from left to right in Fig. 2. The acoustic indicator indices are: 1=Intensity_mean, 2=Intensity_pstd, 3=Intensity_max, 4=Intensity_range, 5=Formant_1_mean, 6=Formant_2_mean, 7=Formant_3_mean, 8=Pitch_mean, 9=Pitch_pstd, 10=Pitch_max, 11=Pitch_range, 12=Duration, 13=SilR.*

Group	Number of people	Common Indicators
A	4	3, 4, 9, 10, 11
B	6	10, 11
C	4	5, 1
D	9	6
E	9	7
F	31	None
G	8	12
H	6	8
I	10	3
J	8	4

scores might implies depression and anxiety so as to lack of rationality. Psychoticism (P): It exists in all individuals with different degrees. Those who score high might love to stay lonely and be inconsideration so as to be hard to accommodate new environment. Lie (L): This factor has not been theoretically specified but has connections with other questionnaires. It represented an stable measurement ability across cultures.

4. Statistical analyses

Before digging into the corpus, we explored how well the subjects performed during the game. We define *successful lie ratio* as the number of successful global lies divided by the number of lies told; *successful lie detection* as the ratio of the number of successful lie detection to the number of lies told; *successful truth detection* as the ratio of the number of successful truth detection to the number of truth told, which is the same as [21]. Results show that, weighted by the number of SUs, the *successful lie ratios* are 56.24% for all subjects, 54.91% for male subjects and 57.52% for female subjects. The SU-weighted *successful lie detection ratios* are 43.76% for all listeners, 34.38% for male listeners and 52.73% for female listeners. The SU-weighted *successful truth detection ratios* are 73.98% for all subjects, 89.88% for male subjects and 60.36% for female subjects. These numbers can be compared against when developing algorithms for automatic deception detection.

Next, it would be of interest to see if there is any correlation between those who successfully deceive others and who successfully detect lies. Results indicate that there is a weak correlation between these two factors ($r(96) = -0.026, p = 0.804$). When considering the gender, however, results indicate that male subjects have positive correlation while female subjects have the opposite (male: $r(48) = -0.144$, $p = 0.330$; female: $r(48) = 0.076$, $p = 0.607$).

The weak correlation of the factors might be due to that the parts each individual put emphasis on were quite different when producing vs. listening to verbal statements. Next, acoustic-prosodic features were individually analyzed for discussion. Section 4.1 describes the definitions of these features. Then t-tests was applied to quantify intra-subject and inter-subject differences. Intra-subject analyses characterizes individual persons and the data seem to suggest that there exists a clustering phenomenon in deceptive behaviours. Inter-subject analyses gives general descriptions for the whole corpus, which helps us locate the significant indicators for automatic deception detection.

Table 2: *The acoustic indicators of deceptive and truthful production/perception. P values are corrected by controlling the FDR at $\alpha = 0.05$. The acoustic indicators (*): $p < 0.05$; non-indicators (-): $p \geq 0.05$.*

Feature	All	Male	Female
Intensity_mean	*/*	-/*	*/-
Intensity_pstd	*/*	*/-	-/*
Intensity_max	*/*	-/*	*/-
Intensity_range	*/*	*/-	-/-
Formant_1_mean	*/-	-/-	-/-
Formant_2_mean	-/*	-/*	*/-
Formant_3_mean	*/*	-/*	*/-
Pitch_mean	*/*	-/*	*/-
Pitch_pstd	*/*	-/*	-/-
Pitch_max	*/*	-/*	*/-
Pitch_range	-/-	-/-	-/-
Duration	-/-	-/-	-/-
SilR	-/*	-/-	-/-

4.1. Feature extraction

In this paper, we focus on the acoustic-prosodic features. Acoustic-prosodic features were extracted using Praat [22] and all feaures were L2-normalized by individual subject to minimize the influences from natural differences. *Pitch*, *Intensity*, *Formants*, *Duration*, *Silence Ratio* and their functionals were selected.

4.2. Intra-subject analysis

There are totally 96 subjects in this corpus with their truthful and deceptive statements individually. Independent two sample t-test was performed on each feature to evaluate if the difference between deceptive and truthful utterances is significant. Here the significant threshold is set at 0.05 as commonly practiced. Based on the p values, a significance-indicator vector $P = [b_1, b_2, ..., b_{13}]^T$ is constructed for every subject, where $i = 1, 2, ..., 13$ is the feature index, and $b_i = 1$ if $p < 0.05$ for feature i and $b_i = 0$ otherwise.

With 96 such P vectors, a dendrogram can be created (Fig. 2) based on the Euclidean distance and the Ward method [23] for linking and clustering. From Fig. 2, we found that there are some common significant features within each small group as listed in Table 1; in other words, each group has a unique set of *acoustic indicators* between lying and truth. However, not all of the subjects have acoustic indicators. The zero-flat part of Fig. 2 (F group in Table 1) shows that the members in this group have no significant acoustic indicator when telling a lie vs. telling the truth.

4.3. Inter-subject analysis

We follow the work in [18] to figure out the acoustic indicators. All of the p values here are corrected by controlling the false discovery rate (FDR) at $\alpha = 0.05$. We compared all truthful utterances with all deceptive utterances and the results are shown

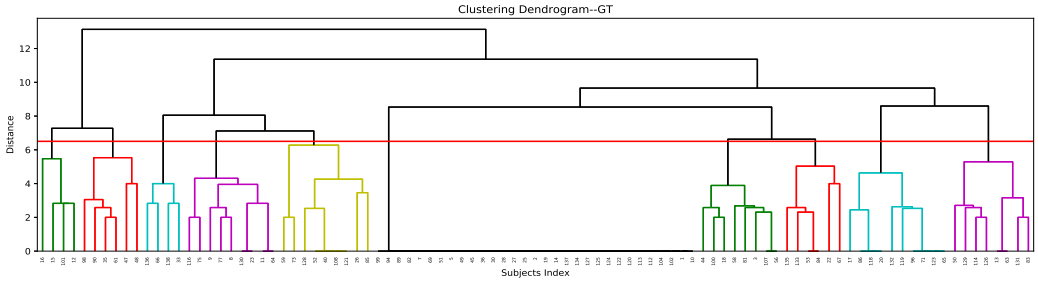

Figure 2: *A clustering dendrogram for intra-subject analysis. X-axis is the subject index and y-axis is the distance between P vectors defined in 4.2. The red horizontal line intersects with 10 branches and thus separates the subjects into groups A to J in Table 1.*

Table 3: *Automatic deception detection of SUs with different demographic groups. Here we show the performance of the Random forest classifier, trained on all features vs. significant indicators of each demographic group described in Table 2.*

Gender	Feature	A	P	R	F1
All	All	53.87	55.36	54.87	53.35
	Sig.	53.93	55.58	55.05	53.37
Male	All	52.28	51.21	51.51	50.58
	Sig.	52.84	52.30	52.58	51.74
Female	All	52.18	51.74	51.69	49.71
	Sig.	52.29	52.42	52.05	49.39

in Table 2. The symbol at the left of slash('/') is from the ground truth of the production of interviewees while the right is from the perspectives of their opponents after hearing the descriptions.

5. Automatic deception detection

An important application related to deception corpus is automatic deception detection [24]. Previous researches mentioned in Section 2 paid much effort on it but a few of them is Mandarin DDD corpus. We used two machine learning algorithms — SVM and Random Forest — to train on all features vs. significant indicators only and validated the model with different demographic groups (e.g. classifying only male or female subjects). Both approaches were implemented using Scikit-learn [25]. For fair evaluation, we separated our corpus into 5 folds depending on the demographic group being investigated. (e.g. if now we focus on the female subjects, we only divide female subjects into 5 folds for training and testing). Then cross training was performed and the average results of testing over these 5 folds are reported in Table 3. Only the results of Random Forest classifier are shown, which turns out to be the better classifier here. We can see that the significant indicators are useful over all subjects except female subjects. We thought it was because the variance between each fold was too large. Some folds could actually achieve approximately 57% of F1 score but some were about 40%. We are still figuring out the problems and hope to add more features in the future. Nevertheless, it still suggests that these intuitive acoustic indicators contain information that may be crucial for deception classification when people make decisions.

6. Discussion and Conclusions

Acoustic-prosodic features chosen here have been proved to possess certain influences when telling and judging lies [7, 26, 27, 28, 29, 30, 31]. It is interesting that the deception behaviour could be clustered in the corpus. The middle flat part of Fig. 2 shows that there certainly are some people who can talk without acoustical differences between deception and truth-telling. From Table 2, the second and third formants are not indicators for male subjects, but results are opposite for female subjects in the side of production. We speculate that it is because most of male subjects speak with a smoother tone and in a slower speaking rate comparing with the female subjects, which might lead to the distinctively different formants for female subjects when telling lies. It is interesting when we look at the both sides at the same time; the acoustic indicators are almost opposite across genders. Another thing that surprised us is that silence ratio is not an indicator for deception production. Perhaps, the subjects not only hesitated before telling lies, but also paused before truthful utterances perhaps because they wanted to prepare credible descriptions to convince their opponents.

We constructed the DDD Mandarin deception corpus and described the details of the procedure. The clustering phenomenon that emerged from intra-subject analysis implies that the subjects could be categorized into groups of common deceptive acoustic features. T-test from inter-subject analysis identifies the significant features and points out the group differences between male and female subjects. Then the automatic deception detection was done and the best classifier achieved the accuracy and F1-score of 53.93% and 53.37%. A meta-analysis of individual differences in detecting deception for students group was provided in [32] and got an accuracy of 54.22%. It shows that deception detection is hard even for human and perhaps harder when only acoustic-prosodic features are available for consideration.

Since deception in daily dialogues is a human-to-human behaviour, efforts have been made to collect both the question and the answer sides of the sessions. The scores from EPQ are also available now. Follow-up studies may consider all the information and combine with linguistic and lexical cues for deception analysis [33, 34]. Hopefully, this corpus can contribute to the advance of automatic deception detection in Mandarin in the future.

7. Acknowledgements

Thanks to National Tsing Hua University for supporting this project.

8. References

[1] E. W. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision support systems*, vol. 50, no. 3, pp. 559–569, 2011.

[2] S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, and A. Rosenberg, "Cross-cultural production and detection of deception from speech," in *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 2015, pp. 1–8.

[3] B. W. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. C. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language." in *Proceedings INTERSPEECH 2016*. ISCA, 2016, pp. 2001–2005.

[4] V. Rubin, "Pragmatic and cultural considerations for deception detection in asian languages," *ACM Transactions on Asian Language Information Processing*, vol. 13, 06 2014.

[5] C. Fan, H. Zhao, X. Chen, X. Fan, and S. Chen, "Construction of chinese deceptive speech detection corpus," in *2015 Joint International Mechanical, Electronic and Information Technology Conference (JIMET-15)*. Atlantis Press, 2015/12. [Online]. Available: https://doi.org/10.2991/jimet-15.2015.18

[6] C. Fan, H. Zhao, X. Chen, X. Fan, , and S. Chen, "Distinguishing deception from non-deception in chinese speech," in *2015 Sixth International Conference on Intelligent Control and Information Processing (ICICIP)*. IEEE, 2015, pp. 268–273.

[7] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception." *Psychological bulletin*, vol. 129, no. 1, p. 74, 2003.

[8] J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. L. Pellom, E. Shriberg, and A. Stolcke, "Distinguishing deceptive from non-deceptive speech," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, 2005, pp. 1833–1836.

[9] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[12] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[13] S. I. Levitan, G. An, M. Ma, R. Levitan, A. Rosenberg, and J. Hirschberg, "Combining acoustic-prosodic, lexical, and phonotactic features for automatic deception detection," in *INTERSPEECH*, 2016.

[14] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[15] C. Whissell, M. Fournier, R. Pelland, D. Weir, and K. Makarec, "A dictionary of affect in language: Iv. reliability, validity, and applications," *Perceptual and Motor Skills*, vol. 62, no. 3, pp. 875–888, 1986.

[16] S. Sondhi, R. Vijay, M. Khan, and A. K. Salhan, "Voice analysis for detection of deception," in *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*. IEEE, 2016, pp. 1–6.

[17] X. Fan, H. Zhao, X. Chen, C. Fan, and S. Chen, "Deceptive speech detection based on sparse representation," in *2016 IEEE 12th International Colloquium on Signal Processing & Its Applications (CSPA)*. IEEE, 2016, pp. 7–11.

[18] S. I. Levitan, A. Maredia, and J. Hirschberg, "Acoustic-prosodic indicators of deception and trust in interview dialogues," *Proc. Interspeech 2018*, pp. 416–420, 2018.

[19] V. Ivkovic, V. Vitart, I. Rudan, B. Janicijevic, N. Smolej-Narancic, T. Skaric-Juric, M. Barbalic, O. Polasek, I. Kolcic, Z. Biloglav *et al.*, "The eysenck personality factors: Psychometric structure, reliability, heritability and phenotypic and genetic correlations with psychological distress in an isolated croatian population," *Personality and Individual Differences*, vol. 42, no. 1, pp. 123–133, 2007.

[20] C. Zhong-geng, "Item analysis of eysenck personality questionnaire tested in beijing-district," *Acta Psychologica Sinica*, vol. 2, 1983.

[21] S. I. Levitan, M. Levine, J. Hirschberg, N. Cestero, G. An, and A. Rosenberg, "Individual differences in deception and deception detection," in *Proceedings of Cognitive*, 2015.

[22] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[23] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[24] S. Sarkadi, "Deception," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 5781–5782.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[26] F. Enos, E. Shriberg, M. Graciarena, J. Hirschberg, and A. Stolcke, "Detecting deception using critical segments," in *INTERSPEECH*, 2007.

[27] L. A. Streeter, R. M. Krauss, V. Geller, C. Olson, and W. Apple, "Pitch changes during attempted deception." *Journal of personality and social psychology*, vol. 35, no. 5, p. 345, 1977.

[28] L. Anolli and R. Ciceri, "The voice of deception: Vocal strategies of naive and able liars," *Journal of Nonverbal Behavior*, vol. 21, no. 4, pp. 259–284, 1997.

[29] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193. Amsterdam, 1993, pp. 97–110.

[30] S. L. Sporer and B. Schwandt, "Paraverbal indicators of deception: A meta-analytic synthesis," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 20, no. 4, pp. 421–446, 2006.

[31] S. Benus, F. Enos, J. B. Hirschberg, and E. Shriberg, "Pauses in deceptive speech," in *Third International Conference of the International Speech Communication Association*, 2006.

[32] M. G. Aamodt and H. Custer, "Who can best catch a liar?" *Forensic Examiner*, vol. 15, no. 1, 2006.

[33] S. I. Levitan, A. Maredia, and J. Hirschberg, "Linguistic cues to deception and perceived deception in interview dialogues," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, vol. 1, 2018, pp. 1941–1950.

[34] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.