INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

The Sound of Deception - What Makes a Speaker Credible?

Anne Schröder, Simon Stone, Peter Birkholz

Institute of Acoustics and Speech Communication, Technische Universität Dresden
anne.a.schroeder@web.de, simon.stone@tu-dresden.de, peter.birkholz@tu-dresden.de

Abstract

The detection of deception in human speech is a difficult task but can be performed above chance level by human listeners even when only audio data is provided. Still, it is highly contested, which speech features could be used to help identify lies. In this study, we examined a set of phonetic and paralinguistic cues and their influence on the credibility of speech using an analysis-by-synthesis approach. 33 linguistically neutral utterances with different manipulated cues (unfilled pauses, phonation type, higher speech rate, tremolo and raised F0) were synthesized using articulatory synthesis. These utterances were presented to 50 subjects who were asked to choose the more credible utterance. From those choices, a credibility score was calculated for each cue. The results show a significant increase in credibility when a tremolo is inserted or the breathiness is increased, and a decrease in credibility when a pause is inserted or the F0 is raised. Other cues also had a significant, but less pronounced influence on the credibility while some only showed trends. In summary, the study showed that the credibility of a factually unverifiable utterance is in parts controlled by the presented paralinguistic cues.

Index Terms: speech perception, articulatory synthesis, computational paralinguistics

1. Introduction

In situated interaction, the credibility of an utterance or of a speaker in general always plays an important role. It impacts the relationship between speakers in a dialog situation, alters the course of conversations, can build bonds of trust or destroy them. But what is it exactly that makes us trust an utterance or, conversely, that raises our suspicions?

When we split utterances into true and false utterances we need to clarify truth and deception. There can be false statements, which are not intended to mislead the listener. In this work, we are talking about statements, which are meant to mislead. In such situations, an influence on the deceiving speaker's speech behavior can be expected because the intentional lie means a higher cognitive load on the speaker and is therefore likely to influence the speaker's mental state (at least, in mentally healthy individuals). It is part of daily life to encounter deception attempts and to identify them as such [1]. Most of these daily lies are harmless and can easily be detected, mostly because there is clear evidence and/or because the interlocutors know each other well, know the history leading up to the situation, and can tell slightest differences in their friend's or family member's behavior. But what about out-of-context situations, like analyzing suspects in criminal investigations, when there is no physical evidence, no provable facts, no background information and the suspect is personally unknown to the assessor? The rates of recognizing whether someone is telling the truth or not vary from 50 % to 61 % depending on whether the listener should pick the false or the true utterance, indicating that it is easier to find true utterances than false ones [2]. A skeptical assessor, who expects to be deceived, can also detect lies more effectively than an unsuspecting listener [3, 4]. But if we have no background knowledge and no means of testing the factual veracity of a statement, how can we detect lies? Since the rates of detecting lies correctly are much better from audio recordings alone than from video-only recordings (with $p < 0.0001$ and an α of 5 % [2]), there appears to be a great deal of information in the *way* a speaker says a lie that may already be enough of a tell-tale. In [5], three major models to explain and predict those paralinguistic features are presented, which are summarized in Table 1.

Table 1: *Theoretical frameworks of deception in speech and their influence on vocal features (recreated from [5]).*

	Arousal theory	**Cognitive theory**	**Attempted control theory**
Acoustic feature	Psychological stress, emotional arousal	Cognitive load	Hyper-articulation
Mean F0	↗	↗, ↘	↗
Mean energy	↗	↗, ↘	↗
Speaking rate	↗	↗, ↘	↗, ↘
Formants	↘	?	↗
Hesitations	↗	↗	↘
Speech errors	↗	↗	↘
Pauses	↘	↗	↗
Phonation type	tense	?	?

↘ = decrease, ↗ = increase, ? = not investigated

The problem is that these predictions are contradictory across, as well as within theories. This is likely because due to the methods used in the studies on deception so far: As can be seen in, e.g., [2, 6–11], there are a number of studies who investigated artificial situations in laboratory settings. While such a setting has the advantage of a controlled environment, it is questionable if those results can be translated to real-life situations. Other studies attempted to avoid artificial settings by using real-word recordings, e.g., from police interviews [6, 8], but in those cases the speech material is very varied, the quality can be shaky, and there can be numerous (emotional and psychological) confounding factors that make the results from these studies almost impossible to compare.

http://dx.doi.org/10.21437/Interspeech.2017-384

So how to resolve this conundrum? In this paper, we propose to turn the paradigm upside-down and instead of *analyzing* deceptive speech, we try to *synthesize* deceptive utterances by systematically manipulating individual (paralinguistic) speech features and examine at which point human raters judge an utterance to be credible or not credible. This process of analysis-by-synthesis is well-established in engineering and was previously employed in vocal emotion research [12, 13] but never in the context of deception.

2. Method

To facilitate the further discussion, we define the following terminology: a *cue* is a particular phonetic or paralinguistic feature manipulated in a certain way, an *item* is a synthesized sequence of digits containing one or no cue, and a *trial* is a forced choice between an item with a cue and a neutral (no cue) item.
A vast number of cues have been examined in the past but, as [14] have shown in their review, not all of these cues were consistently significant when related to the truth content of an utterance. From their exhaustive list, we therefore selected the following cues that had a significant correlation to the veracity of an utterance or at least showed a strong trend (marked by *):

- unfilled pauses
- phonation type*
- higher speech rate*
- tremolo
- higher pitch*

Phonation type and tremolo were not specifically called out as cues in [14], but are likely to be vocal expressions of nervousness and tenseness during attempted deception [5, 14] and were therefore included in this study. To further examine their influence, each cue was used in up to four different variations (see Table 2).

2.1. Synthesis and resynthesis

The items for the experiment were synthesized using articulatory synthesis provided by the free software *VocalTractLab 2.1* [15] developed at our institute. To avoid interference from the linguistic content, the cues were inserted into utterances consisting of sequences of digits (shown in Table 3). Patterns like repetitions or consecutive sequences of digits were avoided and the cues were inserted into digits with comparatively long voiced sections to ensure that they are noticeable. Articulatory synthesis, while being a powerful tool capable to produce the phonetic and paralinguistic manipulations of interest in this study, is a time-consuming method because the entire trajectories of the articulators have to be modeled by hand for each utterance. Since the items in this study all consisted of combinations of digits, it was only necessary to synthesize each of the digits, which could then be concatenated to form any combination. Articulatory synthesis was still preferred over concatenative synthesis to preserve the naturalness of the speech signal when the cues are manipulated. With *VocalTractLab 2.1*, it was possible to manipulate the phonetic features as close to human vocal tract behavior as possible, because the synthesis of this program is based on the aerodynamic and acoustic simulation of speech production based on models of the human vocal tract.

The most straight-forward approach to obtain a natural sounding, neutral utterance is to resynthesize a recording of that utterance by a natural speaker. To that end, a professional speaker - male, 25 years old - recorded the digits zero to nine in German under professional recording conditions in a studio. To ensure a flat intonation of the original words and thus facilitate phonetic manipulations at a later stage, the speaker recorded them embedded in the carrier sentence "Ich habe [digit] gesagt." [ʔɪç ˈhaːbə [digit] gəˈzaːkt(ʰ).]. These words were resynthesized and manipulated with *VocalTractLab 2.1* as detailed in [12], and finalized with the same program and the audio editor and recorder *Audacity 2.0.0* [16]. To illustrate the process of resynthesis, Figure 1 shows the gestural score of the realization of "Null" (IPA: [nʊl], German for "zero") in the *VocalTractLab*. After the resynthesis and manipulation of the individual digits, the words were set together as sentences. In this study, four sentences were used as carrier material for the 14 cues in Table 3. For the neutral versions of the sentences the last word was lengthened by 20 % (phrase-final lengthening) and the fundamental frequency F0 was decreased by two semi-tones. In the neutral case the phonation type was set to modal and the amplitude to -1 dB. Pauses were set to 150 ms duration according to [17].

Because these carrier sentences consisted of three different

Table 2: *Manipulated features investigated in the study, their labels and the number of comparisons versus neutral. Cues that were empirically determined to be difficult to perceive were presented more often. The sentence index refers to the list of carrier sentences in Table 3.*

Manipulated feature	**Cue label**	**Presentations per subject**	**Sentence index**
F0 of 3rd word raised by 4 st	$F0_{+4st}$	3	1
F0 of 3rd word raised by 6 st	$F0_{+6st}$	2	2
Tremolo of 1 st	$tremolo_{1st}$	2	4
Tremolo of 2 st	$tremolo_{2st}$	2	3
Pause of 300 ms between 2nd and 3rd word	$pause_{short}$	2	2
Pause of 450 ms between 2nd and 3rd word	$pause_{long}$	2	3
Breathy phonation	$breathy$	2	3
Slightly breathy phonation	$slightly\ breathy$	2	4
Slightly pressed phonation	$slightly\ pressed$	4	3 & 4
Pressed phonation	$pressed$	2	1
2nd word 20 % shorter	$rate_{+20\%}$	2	1
2nd word 10 % shorter	$rate_{+10\%}$	4	1 & 2
2nd word 10 % longer	$rate_{-10\%}$	2	2
2nd word 20 % longer	$rate_{-20\%}$	2	3
Total number of manipulated items:		33	

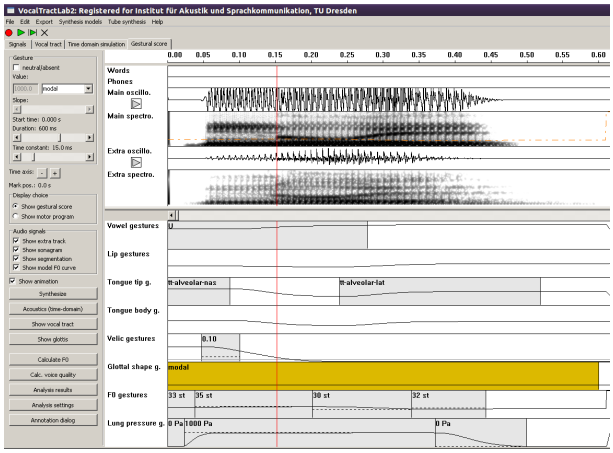

Figure 1: *Gestural score of "Null" in* VocalTractLab. *The signals and spectrograms of the resynthesized and the original version are at the top of the window and the articulatory gestures including glottal shape and F0 modulation are at the bottom.*

Table 3: *The four synthesized carrier sentences and their IPA transcriptions. For the places of manipulation see Table 2.*

Index	German sentence	IPA
1	"Acht Neun Fünf."	[ʔaχt(ʰ) nɔʏn fʏnf.]
2	"Sieben Null Eins."	[ˈziːbən nʊl ʔaɪns.]
3	"Neun Fünf Sieben."	[nɔʏn fʏnf ˈziːbən.]
4	"Fünf Eins Neun."	[fʏnf ʔaɪns nɔʏn.]

digits, three words and two pauses could be manipulated to produce the items for this study. The raised F0 is inserted at the last syllable, the other cues were inserted into the second word. The manipulated pauses were always the second pause in the sentence. The positions were chosen due to their perceptibility and naturalness: a lengthened pause can only be contrasted when compared to a first pause, which establishes the "normal" length. A similar reasoning applied to the position of the other cues, where the first word offers the listener a baseline from which the second word deviates (because of the cue), but the speaker then returns to the baseline for the last word in order to avoid overemphasis of the cue. The raised F0 was realized in the last position, which likens the item more to a question than an utterance and thus might indicate a raised insecurity of the speaker. For variations of the F0, the last syllable was raised by 4 and 6 st. For the tremolo, the voice trembling was set to a 20 Hz oscillation of 1 or 2 st. For the pauses, the second pause was set to 300 ms for short pauses or 450 ms for a variety with longer pause. For phonation type we used the breathy and pressed sound options. The volume was kept constant by adjusting the lung pressure to the phonation type and is therefore higher for the breathy utterances and lower for the pressed phonation type (1200 Pa or 800 Pa, respectively). To decrease or increase the rate of speech, the second digit was lengthened or shortened by 10 or 20 %.

The number of repetitions of each cue was subjectively chosen after an informal listening test: Cues that were difficult to perceive were included more often (see column 3 in Table 2). The total number of manipulated items was 33 thus resulting in 33 comparisons versus neutral.

2.2. Experimental procedure

In the experiment, 50 subjects (32 males, 18 females, age 19 to 52, mean age 28) were presented with 33 trials (see breakdown in Table 2) consisting of a neutral item and an item containing a cue, and one trial consisting of two neutral items, both using a different carrier sentence. The subjects were told that one of the two items was the correct code for a combination lock and the other one was an intentional lie. They were asked to choose the item they think was true and rate their confidence in the decision on a scale from 1 (certain) to 5 (uncertain). The study prompts were presented with the phonetic and sound analysis program PRAAT [18], which also recorded the subjects' responses. Unlimited playback repetitions were allowed. The order of the trials was randomized to avoid systematic familiarization effects.

2.3. Method for evaluation

To assess the credibility that each cue lent to an utterance, the subset of trials containing items with that cue was evaluated. In that subset, each subject was at first examined individually. If a subject always preferred the item with the cue over the neutral item, that cue was awarded a credibility score (C-score) of 1, if the subject always preferred the neutral item, it was scored a -1. If a subject preferred the cue in one trial but not in another, it scored a 0 (thus eliminating inconsistent choices from the total). Once a C-score was determined for each cue and each subject, the final mean C-score was calculated for each cue as the average across all subjects.

3. Results

To prove the aforementioned hypothesis that the use of arbitrary sequences of digits effectively removes the linguistic content as a potential bias, the results from the neutral-vs.-neutral trials were analyzed first. The preference of the subjects of one neutral item over the other was 26 to 24, putting the results well within the 0.5 proportion of the null hypothesis (one sample z-test for proportions, $\alpha = 0.01$). Therefore, we considered any trend or bias towards or against an item versus neutral as the result of the (manipulated) cues.

The results of all the comparisons between manipulated and neutral samples are shown in Figure 2. Each circle represents the mean C-score of a cue across all subjects. The majority of cues (8 out of 14) did not consistently sway the subjects' decisions one way or another, although in many cases trends were visible. The C-scores of the remaining 6 cues, however, significantly deviate from the baseline of 0 (correlating to a neutral credibility) and thus appear to have a consistent influence on the credibility of an utterance.

In particular the raising of the fundamental frequency by 4 st and the long pause caused subjects to refrain from trusting that item and instead trust the neutral one. In contrast, a slight tremolo of 1 st or a slightly breathy phonation type resulted in the subjects trusting those items instead of the neutral ones. To examine if those trends were consistent, we devised another study where the same 50 subjects had to choose between two items which both contained cues, one with a high C-score and one with a low C-score. These results are given in Figure 3. For all four comparisons, the results were consistent with the absolute and the relative results from the first study.

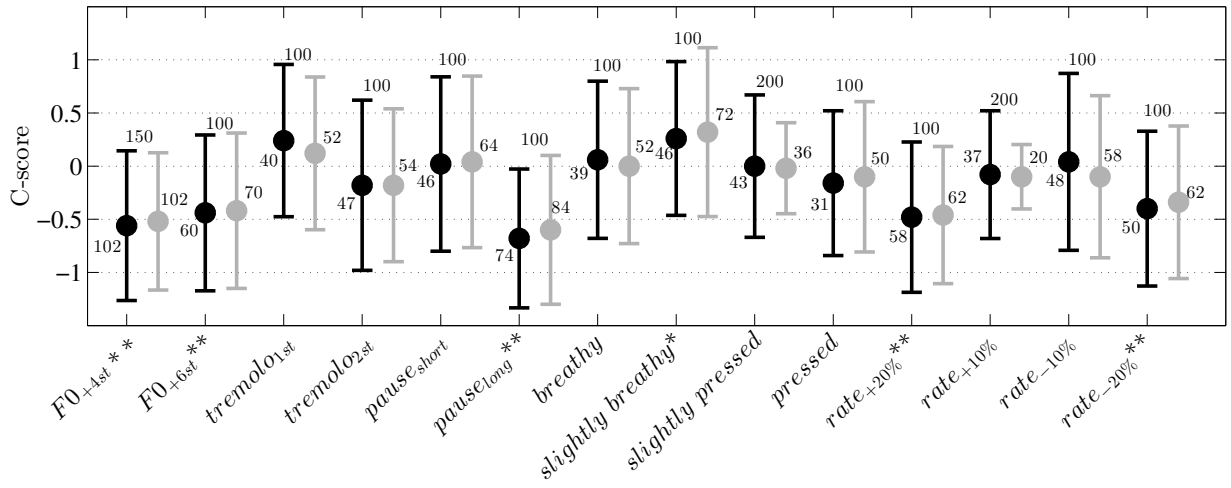

Figure 2: *Credibility score (1: credible, -1: not credible) of all cues when contrasted with a neutral item. Circles are the mean scores across the indicated number of subjects, whiskers indicate the standard deviation. The number above the whiskers is the total number of samples of each cue (perceptively difficult cues were presented more often). Gray: consistent samples (only included when a subject gave the same preference each time they were presented with the respective cue). Black: only samples with a confidence rating of 3 or better were included before checking the consistency. The cues marked by asterisk(s) are significantly different from the indecisive value 0 (two-sample t-test, *$p < 0.05$, **$p < 0.01$).*

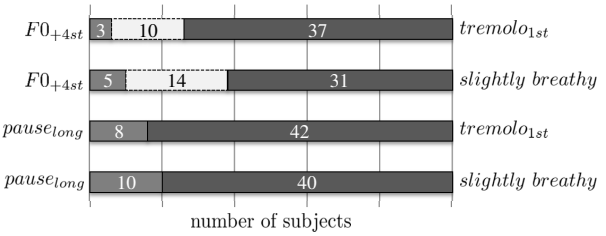

Figure 3: *Results of trials contrasting items with the two highest C-scores (right side) and the two lowest C-scores (left side). The large majority of the 50 subjects preferred the items with high C-scores (dark gray bars) over the items with low C-scores (light gray bars). The white bars show the number of subjects that chose inconsistently across trials.*

4. Discussion

To discuss the results, we recall the basic analysis-by-synthesis idea behind the study: Let the listeners decide, which utterance (with neutral linguistic content) they believe to be true based on the cues. If we look at the cues the subjects apparently used in their decision-making, we find that they are somewhat consistent with the theories about deception in speech: an item with a slightly breathy phonation type was significantly more likely to be trusted by the subjects and an item with a slight tremolo of 1 st also showed a trend towards higher credibility. The tremolo and the slight breathiness may be signs of a relaxed speaking style, which could be a cue for the listener to assume a relaxed, truthful speaker. In contrast, the insertion of pauses appeared to degrade the credibility of an utterance, especially a long pause. These observations are in-line with the cognitive theory, which predicts these changes because of the cognitive load caused by an attempted deception. If, however, the breathiness or the tremolo became too strong, their effect on the credibility became less pronounced and consistent. A too strong tremolo and breathiness may be perceptively indicative of hyperarticulation and interpreted by the listeners as an overreaction of the "speaker" in an attempt to control their speech, as described by the attempted control theory.

5. Conclusions

One interesting observation from the presented study is that numbers made up of interchangeable digits are well-suited stimuli to separate linguistic from paralinguistic information. The results also showed significant influences on the decision-making of the subjects by some of the investigated cues when compared with neutral items. Diametric influences of cues also were consisted when those cues were directly contrasted against each other. In summary, the analysis-by-synthesis approach was successful in this investigation and should be considered for other paralinguistic studies as well (e.g., emotions, sarcasm, intention etc.). The results from this study could be used to place landmarks for further experiments using natural speech material and guide the analysis of speech databases. The results could also be used to make synthetic voices in digital assistants (e.g., Apple Siri, Google Home, Amazon Echo) more credible and thus potentially lower the users' reluctance to talk to a machine. A major limitation of the work so far was that only a single parameter was modified at the same time. Future work should therefore study how covariation of the credibility-carrying parameters identified in this study changes the outcome or if other parameters, which did not significantly influence the credibility, have an impact if they are manipulated together with others.

6. Acknowledgements

The authors would like to thank all subjects for their participation in the study.

7. References

[1] H. S. Park, T. R. Levine, S. A. McCornack, K. Morrison, and M. Ferrara, "How People Really Detect Lies," *Communication Monographs*, vol. 69, no. 2, pp. 144–157, 2002.

[2] J. C. F. Bond and B. M. DePaulo, "Accuracy of Deception Judgments," *Personality and Social Psychology Review*, vol. 10, no. 3, pp. 214–234, 2006.

[3] J. Hilton, S. Fein, and D. Miller, "Suspicion and Dispositional Inference," *Personality and Social Psychology Bulletin*, vol. 19, no. 5, pp. 501–512, 1993.

[4] H. Hörmann, *Psychologie der Sprache*, 2nd ed. Berlin [u. a.]: Springer, 1977.

[5] C. Kirchhübel and D. M. Howard, "Detecting suspicious behaviour using speech: Acoustic correlates of deceptive speech - An exploratory investigation," *Applied Ergonomics*, vol. 44, no. 1, pp. 694–702, 2013.

[6] P. Ekman, M. O'Sullivan, W. V. Friesen, and K. R. Scherer, "Invited article: face, voice, and body in detecting deceit," *Journal of Nonverbal Behavior*, vol. 15, no. 2, pp. 125–135, 1991.

[7] V. L. Cestaro and A. B. Dollins, "An Analysis to Voice Responses for the Detection of Deception," *Department of Defense - Polygraph Institute*, 1994.

[8] K. R. Scherer, S. Feldstein, R. N. Bond, and R. Rosenthal, "Vocal Cues to Deception: A Comparative Channel Approach," *Journal of Psycholinguistic Research*, vol. 14, no. 4, pp. 409–425, 1985.

[9] 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, pp. 345–350, 1977.

[10] B. M. DePaulo and R. Rosenthal, "Telling lies," *Journal of Personality and Social Psychology*, vol. 37, no. 10, pp. 1713–1722, 1979.

[11] 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.

[12] P. Birkholz, L. Martin, Y. Xu, S. Scherbaum, and C. Neuschaefer-Rube, "Manipulation of the Prosodic Features of Vocal Tract Length, Nasality and Articulatory Precision Using Articulatory Synthesis," *Computer Speech & Language*, pp. 116–127, 2016.

[13] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[14] B. M. DePaulo, B. E. Malone, J. J. Lindsay, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to Deception," *Psychological Bulletin*, vol. 129, no. 1, pp. 74–118, 2003.

[15] Peter Birkholz, "VocalTractLab (Version 2.1) [Computer program]," http://www.vocaltractlab.de/, 2013.

[16] Audacity Team, "Audacity (version 2.0.0) [computer program]," http://audacityteam.org/, 2008.

[17] J. Trouvain and B. Möbius, "Individuelle Ausprägung von Atmungspausen in der Mutter- und in der Fremdsprache als Anzeichen kognitiver Belastung," *Studientexte zur Sprachkommunikation*, vol. 71, pp. 177–183, 2014.

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 5.3.52) [Computer program]," http://www.praat.org/, Retrieved March 10, 2017.