JoSS Journal of Speech Sciences

**DOI**: 10.20396/joss.v14i00.20738

# A QUALITATIVE SYSTEMATIC REVIEW OF INTRA-SPEAKER VARIATION IN THE HUMAN VOICE

**SAN SEGUNDO, Eugenia**[1*]
**DELGADO, Jonathan**[2]

[1]Spanish National Research Council
[2]La Laguna University

**Abstract:** *Audio deepfake detection is essential for addressing societal challenges such as differentiating real news from fake content or authenticating voice recordings in legal contexts. However, identifying whether a voice is human or AI-generated requires knowing which characteristics to examine, and the choice of voice features for this task is relatively unguided. This justifies the systematic review presented in this paper. Hypothesizing that human voices exhibit more intra-speaker variation than deepfakes, the aim of this review has been to summarize and analyze the published studies on the topic of intra-speaker variation in human voice. A systematic search was conducted in Web of Science, the Cochrane Library, and the electronic database of the International Journal of Speech Language and the Law, initially identifying 305 studies. After removing duplicates and applying inclusion/exclusion criteria, 36 articles were selected for analysis. Findings highlight speaking style as a major factor in intra-speaker variation affecting various acoustic parameters. This review suggests that experts may prioritize features that show higher within-speaker variation, while noting that their utility for deepfake detection must be verified on deepfake datasets.*

**Keywords:** deepfake; intra-speaker variation; acoustics; speech; artificial intelligence



[1*] Corresponding author: eugenia.sansegundo@csic.es

## 1 Introduction

Recent studies claim that "there is good reason to believe that AI-generated voices will soon be indistinguishable from real ones" (1, p. 9). Testing how true that statement can turn is what has drawn many researchers into comparing voices generated with Artificial Intelligence (AI) and human voices, with the aim of finding a voice parameter (or a set of voice characteristics) that allow us to dismiss such doomed statements, especially since human-like AI-generated voices — namely, deepfakes— have lately been used in a range of crimes and offences (2–4).

A *deepfake* is a synthetic voice generated from deep learning models, particularly neural networks, that bears an extreme resemblance to a real voice and can therefore be used to clone voices and impersonate a speaker (5). Voice biometric systems are vulnerable to this type of technological developments, called *spoofing attacks*, which also seriously endanger the use of recordings as forensic evidence in a legal context, due to the difficulty of distinguishing a real voice from a deepfake (6). Likewise, deepfakes are being used to defame public figures and make them utter false messages, for example in order to influence elections or political decisions.[1] Thus, it is becoming increasingly difficult to distinguish real news from fake news, with the consequence of an unprecedented lack of trust in the media. Therefore, it is of utmost importance to implement a methodology that identifies how to distinguish which voice samples are deepfakes (8).

Indeed, this misuse of technological advances has triggered an important change in a particular discipline: Forensic Phonetics. This is a subdiscipline of Applied Linguistics which applies phonetic knowledge to any type of legal problem involving speech recordings or voice analyses, from the design of voice parades to forensic voice comparison (FVC), i.e., the comparison of an unknown voice, belonging to an offender, with one or several known voices, belonging to the suspect/s (9). So far, the main challenges that forensic phoneticians faced were transmission channel mismatch, potential use of voice disguise in recordings, or else the low quality (e.g., telephone-degraded, background noise, etc.) and/or short duration of voice recordings in real casework (10). The biggest issue nowadays has turned to whether a disputed recording is real (i.e., produced by a human) or has been generated artificially. However, it is unclear what voice parameters experts should observe and analyze to answer this question.

The choice of voice characteristics to solve the question of deepfake detection in experimental studies by speech researchers so far is relatively unguided (11, 12), in the sense that either studies are too preliminary (case studies; i.e., one speaker analyzed in his original natural voice and in his AI-generated counterpart) or studies analyze only a few parameters, chosen on the basis that they are known to work well for distinguishing speakers (that is, human speaker comparison). However, there is a lack of studies that provide a list of voice parameters to which experts should pay attention to in order to answer the question of whether a voice is of human origin or created with AI techniques. This justifies the systematic review carried out in this paper.

Our aim is not to survey the current state of research in voice presentation attack detection (PAD) or voice spoofing attacks, which are the names given in the automatic speaker verification (ASV) literature to the kind of attacks launched by a malicious party in order to impersonate an authorized individual in a voice recognition system and unlawfully bypass it (13). For that purpose, previous studies exist. For instance, Tan et al. (13) surveyed 172 articles published

[1] As a case in point, in the 2023 Slovak parliamentary elections, deepfake audio recordings were circulated online just two days before the vote, featuring fabricated voices of political candidates discussing plans to manipulate the election results. Despite being quickly debunked, the recordings spread widely on social media and messaging platforms, raising concerns about the role of generative AI in undermining democratic processes and influencing voter behavior (7).

between 2015 and 2021 with the aim of systematically analyzing the state-of-the-art in voice PAD systems, providing a useful taxonomy of types of spoofing attacks, as well as highlighting common issues and future directions of work. Acknowledging the importance of conducting such surveys for the Computer Science community, we still find an important research gap in this area in terms of the phonetic-acoustic aspects that should be considered for voice deepfake detection. It is taken for granted that the sort of features that perform optimally in tasks such as speaker recognition will perform equally well in voice PAD systems too. Even though Tan et al. (13) indicate that using multiple features is more common than using a single feature, popular features seem to be limited to MFCC (i.e., the coefficients that make up a Mel-Frequency Cepstrum, MFC), or CQCC (i.e., the coefficients extracted from Constant Q Transform, CQT), with different classifiers; namely, conventional, deep learning and multiple classifiers. These are calculated from frames or windows trying to describe the spectral envelope of sound, as a way to account for the resonance properties of the vocal tract. However, there is a lack of mapping between those coefficients and the nuances of voice sounds that are typically investigated in Phonetics: pitch, vowel formants, articulation rate, voice quality, etc.

In contrast to previous systematic reviews, our investigation departs from a clear hypothesis, which is that human voices present, as a common denominator, a large intra-speaker variation, which may not be so large in deepfakes. If this hypothesis proves true experimentally, when analyzing a potential deepfake, experts should pay attention to phonetic features that are known to exhibit large intra-speaker variation. The first step is to know which phonetic features show large intra-speaker variation and in which circumstances (language, speaking style, etc.). For that purpose, our systematic review has examined previous phonetic studies that have delved into intra-speaker variation in any voice and speech aspect (see section 2: Methods).

Also called within-speaker variation, intra-speaker variation commonly refers to variation within the same speaker, either from one recording to another, or variation of a voice characteristic in one speaker within the same recording. In FVC, the fact that a voice parameter presents high intra-speaker variation is undesirable. Ever since the first pioneer works in this discipline appeared (14–16), forensic experts have agreed on the existence of a few criteria for selecting an optimal parameter for FVC, insisting on the fact that voice characteristics should be as consistent as possible for each speaker (low intra-speaker variation). In other words, optimal parameters are those which show consistency throughout the utterances of an individual (15). Previous forensic investigations (e.g., 17, 18) reflect the quest for an optimal voice parameter that can distinguish different speakers while remaining stable within a speaker. While some studies have found features that remain constant within a speaker, as we will discuss in this paper, other studies have found the opposite (16). The latter precisely back and endorse our hypothesis; namely, that intra-speaker variation is a common human feature. Most importantly, we hypothesize that such variation is either absent or significantly reduced in synthetically generated (fake) voices. In a very recent study, Ross, Corley and Lai (19) provide some evidence for a particular parameter: pitch range, which refers to the difference between the lowest and the highest fundamental frequency (f0) values produced in an utterance. Their study found a smaller pitch range for deepfakes than for real voices. Larger variation in pitch range can occur in humans due to a number of reasons: emotional, prosodic, health-related issues, etc. Since this variation is difficult to control voluntarily, it makes part of what makes us 'human'. How many other parameters would show this large intra-speaker variation?

As the question of 'what makes a voice fake and what makes a voice human' is largely under-researched, it seems very timely to draw on the knowledge derived from phonetic studies in order to systematically review which voice parameters show large intra-speaker variation.

Intra-speaker variation is here defined following Bell (20), as "the range of variation for particular sociolinguistic variables produced by individual speakers within their own speech" (20, p. 90). This concept is different from intraindividual variation, which is a broader concept. According to Bülow and Pfenninger (21), intraindividual variation refers to any observable variation within individuals' behavior. The notion of intra-speaker variation is used in variationist sociolinguistics not only to refer to systematic individual variability in the context of style and discourse (22, 23), but also to individual language behavior. As highlighted by Bülow and Pfenninger (21), Labovian variationism regards individual speakers as representatives of a speech community (e.g., members of a social class or ethnic group) but minimal attention has been given to speaker-inherent variations that may occur independently of the context or communication partner, without following a systematic pattern, or speaker variation over time. Acknowledging that the purpose of this systematic review is to address an applied phonetic topic (human intra-speaker variation in the AI era for deepfake detection), we consider that the results of our review could be of interest to a wider audience of scientists, for instance linguists who have recently discussed the importance of reconciling psycholinguistic and sociolinguistic approaches to intra-individual variation.

## 2. Method

### 2.1. Identification and selection of studies

Two authors independently carried out the literature search using the electronic databases Web of Science, and the Cochrane Library (subtopic: ear, nose and throat > larynx), as well as the electronic database of *International Journal of Speech Language and the Law*, which requires member subscription. They performed a search with the terms listed in Table 1. An effort was made to include synonyms to account for the possible ways to describe the condition being studied: acoustic parameters with large intra-speaker variation. No publication date restrictions were made. Open Access (OA) studies available in the databases specified above were included. Grey OA and black or dark OA studies were excluded. Grey OA refers to research uploaded to a social network or to the researcher's own website. Black or dark OA refers to scientific literature that is made available through shadow libraries and other channels, most of which are illegal.

In- and exclusion criteria are listed in Table 2. Hand search of the reference lists of the included articles was conducted to identify any further articles, but did not yield any additional studies. The most recent search of the literature for this study was conducted in February 2024[2].

**Table 1:** Systematic syntax

| |
|---|
| (intra speaker OR intraspeaker OR intra-speaker) OR (within speaker OR withinspeaker OR within-speaker) AND (variation OR variability) AND (acoustic) AND (forensic) AND (identity OR disguise) OR (human OR machine) |

[2] This review was conducted in accordance with a previously published protocol (63), which defined the research question, eligibility criteria, and analysis procedures prior to data collection.

**Table 2:** In- and exclusion criteria

| | |
|---|---|
| *Inclusion criteria* | Acoustic measures of intra-speaker variability |
| *Exclusion criteria* | Diachronic variation |
| | General introductory studies |
| | Exclusively computational or statistical studies |
| | Perceptual or qualitative studies |

## 2.2. Data extraction (selection and coding)

The review process for data extraction was conducted with the software *Covidence*, a web-based collaboration software platform that streamlines the production of systematic reviews (24). The process comprised two stages. Both stages were carried out by two independent researchers. First, the title and abstract of the search results were screened against the inclusion and exclusion criteria. Those which met the inclusion criteria were moved through to the second stage of reviewing the full article. The two reviewers evaluated each article blindly at ones' own rate by assigning one of these three labels: "yes", "no" or "maybe". If both rated a reference with "yes", this reference was automatically accepted for further consideration. If both rated a reference with "no", this reference was automatically rejected. If one reviewer rated a reference with "yes" and the other with "maybe", this reference was automatically accepted for further evaluation. If one reviewer rated a reference with "yes" and the other with "no" (as well as if one reviewer rated a reference with "maybe" and the other with "no"), this reference moved to the section "conflicts". In those cases, a meeting was held to solve conflicts upon consensus.

The very few differences in rating were settled by consensus agreement after a discussion. Consulting a third reviewer was therefore not necessary.

Data extraction and synthesis followed the thematic synthesis approach (25). Extracted data included the following information in an Excel spreadsheet: study information (study ID, author, journal and year of publication), objectives of the study, method of data collection method, study setting, data analysis method, and the dependent variable of the study: intraspeaker results (i.e., the acoustic results related to intraspeaker variation).

## 2.3. Risk of bias (quality) assessment

The Critical Appraisal Skills Programme (CASP) Qualitative Research Checklist (26) was used to appraise the methodological quality of all papers assessed for eligibility in the screening (see Figure 1). The decision to include or exclude a paper was discussed among the researchers and agreed upon mutually. Two reviewers independently appraise the included studies. There was no disagreement between the reviewers, so consulting a third reviewer was not necessary. Furthermore, an audit trail of evidence of the included and excluded studies that clarify the reasoning was kept.

The Critical Appraisal Criteria that we used (Table 3) includes 9 questions extracted from the CASP checklist (26). All questions in the original list are used, except "is there a qualitative methodology appropriate?", since our reviewed studies included both quantitative and qualitative methodologies. The list is aimed at helping reviewers appraise studies systematically. Three possible replies are possible: "yes", "no" and "can't tell". The questions of the checklist template

available online (https://casp-uk.net/casp-tools-checklists/qualitative-studies-checklist) were introduced and answered in the same Excel spreadsheet in which data extraction was conducted for the 86 studies assessed for eligibility in the screening process.

**Table 3:** Critical Appraisal Criteria adapted from CASP checklist (26)

| *Questions* |
|---|
| 1. Was there a clear statement of the aims of the research? |
| 2. Was the research appropriate to address the aims of the research? |
| 3. Was the recruitment strategy appropriate to the aims of the research? |
| 4. Was the data collected in a way that addressed the research issue? |
| 5. Has the relationship between researcher and participants been adequately considered? |
| 6. Have ethical issues been taken into consideration? |
| 7. Was the data analysis sufficiently rigorous? |
| 8. Is there a clear statement of findings? |
| 9. How valuable is the research? |

# 3. Results

## 3.1. General results

In total, 305 articles were found, as displayed in Figure 1. A first selection was made based on abstract and title by two independent reviewers. Next, the final inclusion was made using the original full-text articles and the in- and exclusion criteria (Table 2). The PRISMA flow chart shown in Figure 1 was automatically generated by the software *Covidence* after finishing the systematic review. The different sections and subsections, as well as the labels used, had not been edited. The reasons for exclusion correspond to the labels used through the screening process. They have been manually rearranged in Figure 1 to show first the most likely reasons for excursion and then the least likely.

Out of 305 found studies, 55 duplicates were automatically identified by *Covidence* and removed. Reviewers proceeded then to review 250 studies, out of which 164 were excluded in the first screening stage (title and abstract). Then, 86 studies were assessed for eligibility: the full text was read by the two reviewers. Finally, 50 studies were excluded for one of the reasons specified in Figure 1. More than half of these studies were discarded because intra-speaker variability was not actually measured. This decision was taken upon reading the full article, even though the first screening (i.e., title and abstract) could lead us to include it. The second most frequent reason was that the study did not include acoustic measures. These two reasons go against the inclusion criterion in Table 2. As for the exclusion reasons, it has to be noted that in a systematic review the exclusion reasons need to be discussed by reviewers in advance, i.e., prior to undertaking the review. There were four exclusion reasons that were agreed upon (see Table 2). Out of those four, that we foresaw as potential reasons to exclude a paper, only two actually occurred: perceptual study (n=4) and exclusively computational and/or statistical study (n=3).

Furthermore, we had to exclude papers for six more reasons that we had not thought of in advance, but were valid reasons that arose during the screening process. Two are rather technical

reasons: one reference was not actually a paper, but a conference abstract; another was discarded because it was written in a Slavic language that we could not read. Only the abstract was in English, but we could not proceed to read the full paper. As for the rest of the reasons, four studies involved the acoustic analysis of sustained vowels, which was completely different to the type of natural language (reading or spontaneous speech) that the rest of papers included. Similarly, another study presented laboratory conditions with scarce ecological validity. Two studies focused on non-adult populations, which was deemed out of the scope of this paper. Finally, another study presented a case study with a single speaker and no statistics. Because it was rather descriptive and no empirical results were offered, it was also discarded.

### 3.2. Quality of included studies

Only one study was discarded because it did not meet one of the Critical Appraisal Criteria. It did not include information about the age of the speakers, so we considered that it could not be score on, at least, these three questions: "Was the recruitment strategy appropriate to the aims of the research?", "Was the data collected in a way that addressed the research issue?" and "Have ethical issues been taken into consideration?"

### 3.3. Results on intra-speaker variation

Table 4 summarizes the data of the included studies on intra-speaker variation. The studies are grouped according to the main type of acoustic parameter or sound that was investigated. Seven groups were identified: a) temporal parameters (9 studies); b) formants (7 studies); fundamental frequency (6 studies); disfluencies (4 studies), fricatives (3 studies); voice quality (3 studies); other/varia (5 studies).

There was only one study that did not fit into one of the set groups because it delved into both temporal parameters and fundamental frequency. So we have included it in both groups. The miscellaneous nature of the five studies making up the group "other/varia" did not allow us to include it in any specific group. Each study investigates a different speech phenomenon, so they do not share any characteristics that allowed us to group them in any different way.

Results are derived from 36 different papers. The papers included are dated from 1976 to 2022, therefore spanning 46 years. The number of speakers in each study ranges from 5 to 543 (mean: 53.17; mode: 10). The studies cover 15 main languages, including different varieties per language (e.g., American English, Australian English and British English).

For the sake of simplification, Table 4 does not include the exact method for acoustic measurement, but there is considerable methodological variability across studies that might affect the results. When discussing the implications of the findings, this aspect should be taken into account. For instance, in (27) speech rate was calculated as the mean of the log pointwise speech rates of all utterances having four or more words, while in (28) speech rates were calculated as syllables per second. In the group of parameters investigating formant frequencies, the main difference lies in whether formant frequencies are analyzed at the midpoint of the vowel (static approach) or considering the temporal trajectory of the formant (dynamic approach typically considered when analyzing diphthongs and triphthongs). Even within the disfluencies parameter group, which focuses on the same language variety and speaking style, there is methodological variability in what has been measured and how.

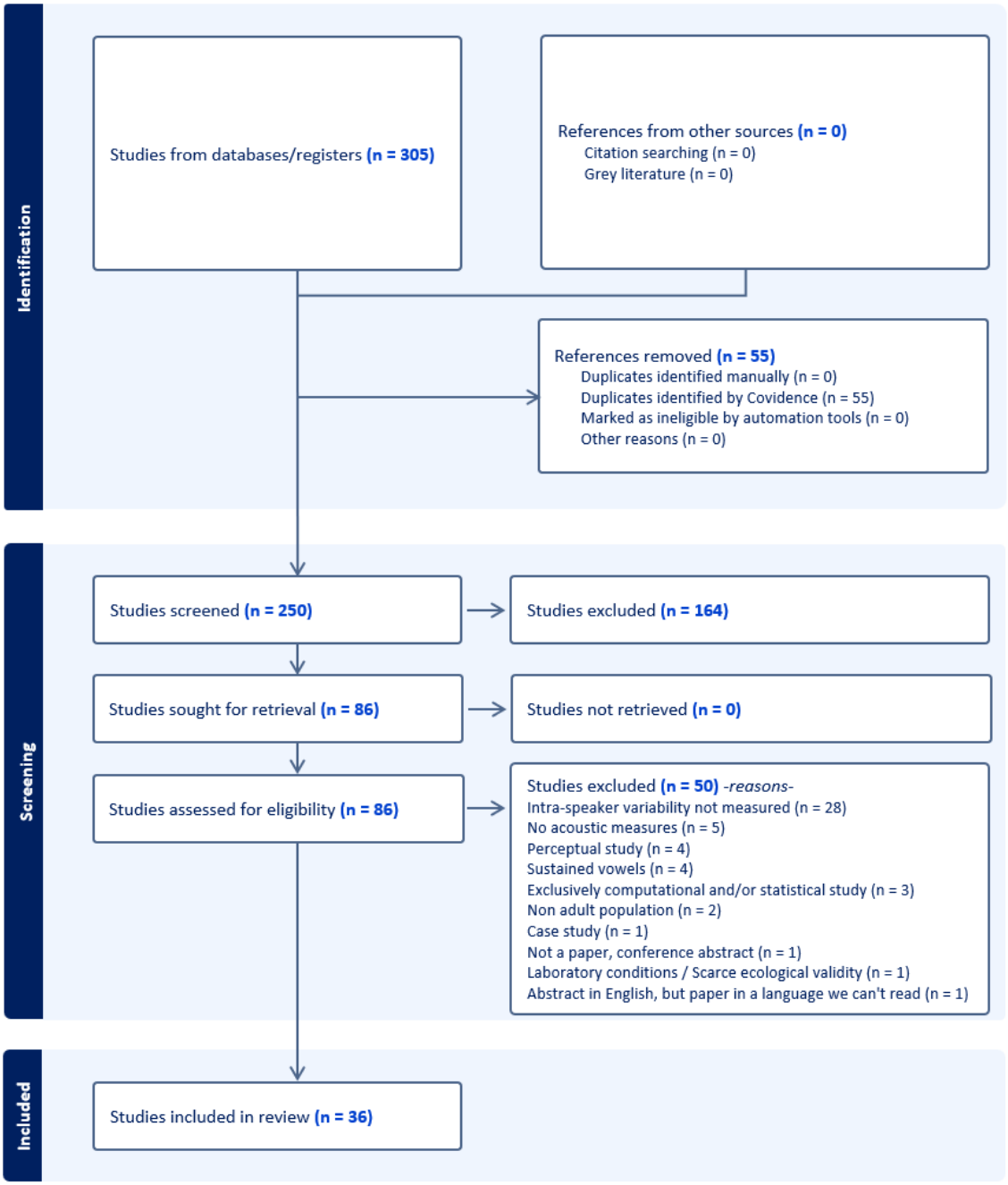


**Figure 1:** PRISMA flow diagram. Results of study selection: references imported from databases, removed, screened, irrelevant, assessed for eligibility, excluded and finally included.

**Table 4:** Intra-speaker variability studies

| *Ref.* | *Type of acoustic parameters* | *N* | *Language* | *Speaking style* | *Acoustics measures* | *Authors' Conclusion* |
|---|---|---|---|---|---|---|
| (38) | Temporal | 28 | Standard German | Reading and spontaneous speech | Rate of consonant or vocalic intervals per second (rate CV), percentage over which speech is vocalic (%V), and interval variability measures | Speech rhythm is a personal trait that remains constant within speakers. |
| (45) | Temporal | 34 | Persian | Reading speech | Articulation rate, %V and vocalic (ΔV(ln), n-PVI-V) and consonant duration variability measures (ΔC(ln)) | Articulation-rate variability within speakers did not exert a significant influence on %V variability. However, it was observed to have a pronounced impact on the duration of vocalic and consonantal measures. |
| (46) | Temporal | 20 | Brazilian Portuguese | Spontaneous dialogues | Speech rate, articulation rate and syllable, vowel and pause durations | The speech rate and articulation rate exhibited remarkable consistency within speakers during spontaneous speech. |
| (47) | Temporal | 38 | German | Reading speech | Mean and peak intensity variability across syllables | No intra-speaker differences were observed in any of the measures studied. |
| (28) | Temporal | 190 | American English | Reading and spontaneous speech | Articulation rate and reading rate | The rate of articulation in spontaneous speech is faster than that in read speech, and there is a correlation between the two rates: speakers with a faster speaking rate have also a faster reading rate. |
| (39) | Temporal | 16 | Swiss German | Reading and spontaneous speech | Durational variability of consonantal and vocalic intervals, voiced and unvoiced intervals, and syllable-peak-to-syllable-peak intervals | The percentage over which speech is vocalic (%V) and voiced (%VO) show high between-speaker and low within-speaker variability. %V and %VO are robust to speaking style variability. %VO is robust to channel variability. |

| | | | | | | |
|---|---|---|---|---|---|---|
| (27) | Temporal | 543 | American English | Spontaneous | Speech rate | A speaker's speech rate converges with the interlocutor's speech rate, exhibiting an increase or decrease in accordance with the speech rate of the conversational partner. |
| (29) | Temporal | 160 | Dutch | Spontaneous | Articulation rate | Most of the within-speaker variance in tempo can be ascribed to the length of a phrase, due to anticipatory shortening. |
| (42) | Temporal / Fundamental frequency | 22 | Korean-English/ Mandarin-English | Reading and spontaneous speech | Mean fundamental frequency (F0), F0 variability (F0 DS), intensity, and speech rate | Bilingual speakers showed significant differences in the average F0, F0 variability, and speech rate in their two languages. However, intensity did not reveal any language or task effect. |
| (48) | Fundamental frequency | 5 | English | Reading speech | F0 | There is a great deal of variability in F0 between different speakers. There is essentially as much variability within a single speaker as there is between several speakers of the same sex. |
| (40) | Fundamental frequency | 16 | Swiss German | Reading and spontaneous speech | F0 declination, local excursion of F0, duration of local F0 excursion, temporal alignment of F0 rise relative to syllable onset and offset | There is a considerable degree of variability between speakers, as well as a relatively low degree of variability within speakers, with regard to a number of F0 features, when considering the two distinct speaking styles. |
| (49) | Fundamental frequency | 20 | Thai | Reading monosyllabic words in isolation | Average F0 and standard deviation in pitch contours of Thai tones | The degree of intersubject variability exceeded that of intrasubject variability when the data were pooled across tones. Static tones (mid, low, high) exhibit greater variability than dynamic tones (falling, rising). T |
| (41) | Fundamental frequency | 107 | German | Reading and spontaneous speech | F0 and relative standard deviation of F0 (F0varco) | In spontaneous speech, the parameter F0varco demonstrates a high degree of stability with respect to changes in vocal effort. There is considerable variation in F0 between speakers when vocal effort is varied. The discrepancy between read and spontaneous speech is relatively minor. |
| (43) | Fundamental frequency | 100 | German | Reading speech | F0 and F0SD | A consistent correlation exists between the F0 of a speaker's typical speech patterns and the type of disguise they employ in a mock incriminating telephone conversation. Three types of disguises were elicited: high pitch, low pitch and denasal voice. |
| (34) | Formants | 14 | Catalan | Reading speech | Stressed midvowels | Speakers are not always consistent in their realization of midvowels /e/ vs. /ɛ/; /o/vs./ɔ/) when they produce the same word. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | When overlap in F1 space was observed, it was more common and extensive for the posterior midvowels. |
| (17) | Formants | 10 | Chinese | Reading speech | The first four vowel formants (F1, F2, F3 y F4) of monophthongs, diphthongs and triphthongs | For each speaker: less intra-speaker variation than corresponding inter-speaker variations. F3 and F4 have greater F-ratio value than F1 and F2. It supports the idea that greater individual differences are found on F3 and F4. |
| (31) | Formants | 5 | Australian English | Reading speech | F1, F2, and F3 of /aI/ examined at equidistant time-normalized intervals. | Greater intraspeaker differences for nuclear-stressed /aI/ than for non-nuclear /aI/. |
| (33) | Formants | 10 | American English | Clear-speech, citation, reading and conversation | F1, F2, F3 in four corner vowels and four lax vowels, and the Euclidean distances of vowels in the F1–F2 and F2–F3 planes | The data revealed considerable variation in the productions of conversation within and across speakers. The values for F1 and F2 remained relatively consistent across tasks. The clear-speech F3 value exhibited the greatest magnitude, whereas the reading demonstrated the most pronounced reduction. |
| (32) | Formants | 9 | Dutch | Spontaneous conversation | F0, F1, F2, duration, and intensity | The degree of variation within speakers was found to be less pronounced in content words than in function words. |
| (18) | Formants | 8 | British English | Script Speech | F1, F2 and F3 of monophthongs and diphthongs | The frequencies of formants decline with age. The most substantial and pervasive alterations are observed in F1, whereas F3 exhibits the greatest stability. |
| (30) | Formants | 5 | Swedish | Reading speech | F1, F2, F3, F4 in segments vowel+/r/+vowel | The majority of the variation within speakers is observed in F2 and F4. |
| (44) | Disfluencies | 162 | German | Reading and spontaneous speech | Pauses, repetitions, false starts, interruptions and unusual phone lengthening (durations and rate of occurrence) | All disfluency indices exhibit an increase when a notable transition occurs between sobriety and drunkenness, with the exception of false starts and repetitions, which demonstrate a decrease in spontaneous speech. |
| (35) | Disfluencies | 20 | British English | Spontaneous | Pauses, repetitions, prolongations, interruptions (measured as disfluencies per minute) | Some subjects exhibited relatively consistent overall rates during the interview and conversation. In contrast, other subjects exhibited considerably greater variability. |

| | | | | | | |
|---|---|---|---|---|---|---|
| (36) | Disfluencies | 60 | British English | Mock police interview | Midpoint frequencies of F1, F2 and F3 in the vocalic portion for /um/ and /uh/ | There is a considerable degree of variation observed in the phonetic quality of the vocalic portions of both /um/ and /uh/, both within and between speakers. |
| (37) | Disfluencies | 20 | British English | Mock police interview | Silent pauses, filled pauses, repetitions, prolongations, and interruptions (measured as the number of occurrences of each type of phenomenon per 100 syllables of speech) | Individual differences in disfluencies are evident among speakers in both styles (interview and telephone conversation). Correlation analyses showed, for many features, strong patterns of correspondence between individuals' usage of a particular feature across the two styles (i.e., intra-speaker consistency). |
| (50) | Fricatives | 14 | Colombian Spanish | Reading and narration speech | Center of gravity (CoG), skewness and kurtosis of the fricative noise in normative [-s̪t-] and assibilated [-s̪t̪-], as well as frequency of occurrence of each variant. | The role of individual behavior, including the sex of the speaker, plays an important role in phonetic reduction. Idiosyncratic variations are reflected in the frequency of occurrence of normative and assibilated variants. |
| (51) | Fricatives | 10 | American English | Reading speech | Spectral mean and skewness of the fricative noise in alveolar /s/ and palato-alveolar /ʃ/ fricatives | The magnitude of the /s/ - /ʃ/ difference varied across speakers, even within gender. The reason for this variability is unknown, but may be linked to a host of factors, including dialectal, anatomical, or auditory-perceptual inter-speaker differences. |
| (52) | Fricatives | 28 | English | Non-sense words repetitions | Fricative /s/ and /ʃ/ spectral mean | Fricatives /s/ and /ʃ/ were more variable before /p/ than before /t/. |
| (53) | Voice quality | 10 | Belgian French | Reading speech | Long-Term Average Spectrum (LTAS) | The degree of intra-speaker variability is contingent upon the subject in question. Some speakers produce utterances for which the LTAS are highly similar to one another. Conversely, other speakers produce utterances for which the LTAS are weakly similar. |

| | | | | | | |
|---|---|---|---|---|---|---|
| (54) | Voice quality | 20 | German | Reading speech | Sublaryngeal and laryngeal voice quality parameters | Voice quality parameters H1* - A2* and H1 * - A3* are similar to H1 * - A1 and B1 in displaying a considerable degree of intra-speaker variability. With respect to H1* - H2*, inter-speaker differences are more pronounced than intra-speaker differences. |
| (55) | Voice quality | 100 | English | Reading speech | 26 acoustic variables from a psychoacoustic model of voice quality measured every 5ms on vowels and approximants | The greatest acoustic variability observed within individual speakers was attributed to variability in source spectral shape and spectral noise. |
| (56) | Others/Varia | 10 | Non-rhotic Southern British accents | Reading speech | F3 minimum within each selected /r/ | Three of the five speakers whose native /r/ is acoustically and auditorily less standard were clearly able to produce /r/s which have lower F3 and are rated as standard. Vocal disguise intra-speaker variation implications: at least some speakers who usually realize /r/ nonstandardly would be able to affect standard /r/ quite readily in vocal disguise. |
| (57) | Others/Varia | 10 | Jamaican English/ Jamaican Creole | Citation speech, Infant Directed Speech, Hyperspeech (Lombard speech) | F0, F1, F2, segmental duration, and intensity in vowels embedded in an /h_d/ context | Infant-directed speech (IDS) and Lombard speech showed similar adjustments in two parameters (F0 and intensity), and IDS and Citation speech showed the greatest spectral differences. Across tasks, mean segmental durations were greatest in the hyperspeech and citation conditions. Lombard condition tokens were, on average, produced loudest, relative to environmental conditions, of all four speech tasks. |
| (58) | Others/Varia | 34 | German | Reading and spontaneous speech | F0 and Vowel Space Size | A comparison of the Adult-Directed Speech (ADS) of the mother and father reveals intraspeaker variation from before the child was born up to 10 months after birth. This variation is observed at three time points. Mothers exhibited a continuous decline in mean F0, while fathers demonstrated a fluctuating pattern, initially showing a reduction followed by an expansion in vowel space size. |
| (59) | Others/Varia | 50 | Japanese | Reading speech | Acoustic correlates for perceived vowel | Within-speaker stability was found in the degree of nasalization of non-nasal vowels and less stability in syllables with oral onsets. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | nasalization (A1– P1, F1 and Fp1) | |
| (60) | Others/Varia | 6 | German, South Bavarian Tyrolean Dialect | Reading speech | Duration measures and Harmonic-to-Noise Ratio (HNR) for the acoustic characterization of /r/ allophony. | High intra- and inter-speaker variability in the realization of /r/. |

# 4. Discussion

## 4.1. Discussion of overall results

The aim of this systematic review has been to summarize and qualitatively analyze the published studies on the topic of intra-speaker variation in human voice to shed some light on what makes us humans, which can distinguish us from artificially-generated voices. The 36 papers analyzed were classified in seven groups (Table 4), according to the main type of acoustic parameter or groups of parameters that the authors focused on. There are no striking differences between these seven groups in terms of sample size or analyzed language, with two notable exceptions.

On the one hand, there is a marked contrast between sample size in studies investigating temporal parameters and studies investigating formants. The former studies recruit a larger number of speakers, typically more than 16, sometimes quite higher: N = 190, 543 and 160 (27, 28, 29). In contrast, the studies focusing on formant frequencies present a lower sample size. N = 5, 5, 8, 9, 10, 10, 14 (30 31, 18, 32, 17, 33, 34). For the other five parameter groups, we find intermediate sample sizes.

On the other hand, there is a marked contrast in terms of language distribution between the group of studies investigating disfluencies and the rest of the parameter groups. There are only four studies focusing on disfluencies and three of them investigate British English (35, 36, 37). Furthermore, they all use the same corpus, so this should be considered at the time of extrapolating results for this specific group. In all the other parameter groups, there is a balance between analyzed languages, with at least three different language varieties investigated per group.

Speaking style is an important variable in intra-speaker variation research. While most studies (41.17 %) investigate reading speech alone, it is not uncommon within a study to present comparative results between reading and spontaneous style (26.48 % of the analyzed studies do so). Spontaneous speech alone is investigated in 23.53 % of the studies. The remaining 8.82 % of the studies focus on a different style which cannot be considered neither spontaneous nor reading. For instance, it is common to elicit disfluencies from a mock police interview (36, 37), which is a specific type of speaking task in forensic corpora.

Figure 2 shows that voice quality studies are conducted solely on reading speech. In contrast, disfluencies are investigated almost exclusively on spontaneous speech (together with the above-mentioned eliciting task, the mock police interview), which makes sense, since disfluencies refer to regular and irregular interruptions in the natural flow of speech. For the rest of parameter groups, we observe a balance in the type of speaking styles investigated. Temporal parameters are investigated in all the possible speaking styles: reading only, spontaneous only and comparing their performance in both styles. For fundamental frequency and formants, it is common to investigate intra-speaker variation in reading-style, followed by a combination of reading and spontaneous styles.

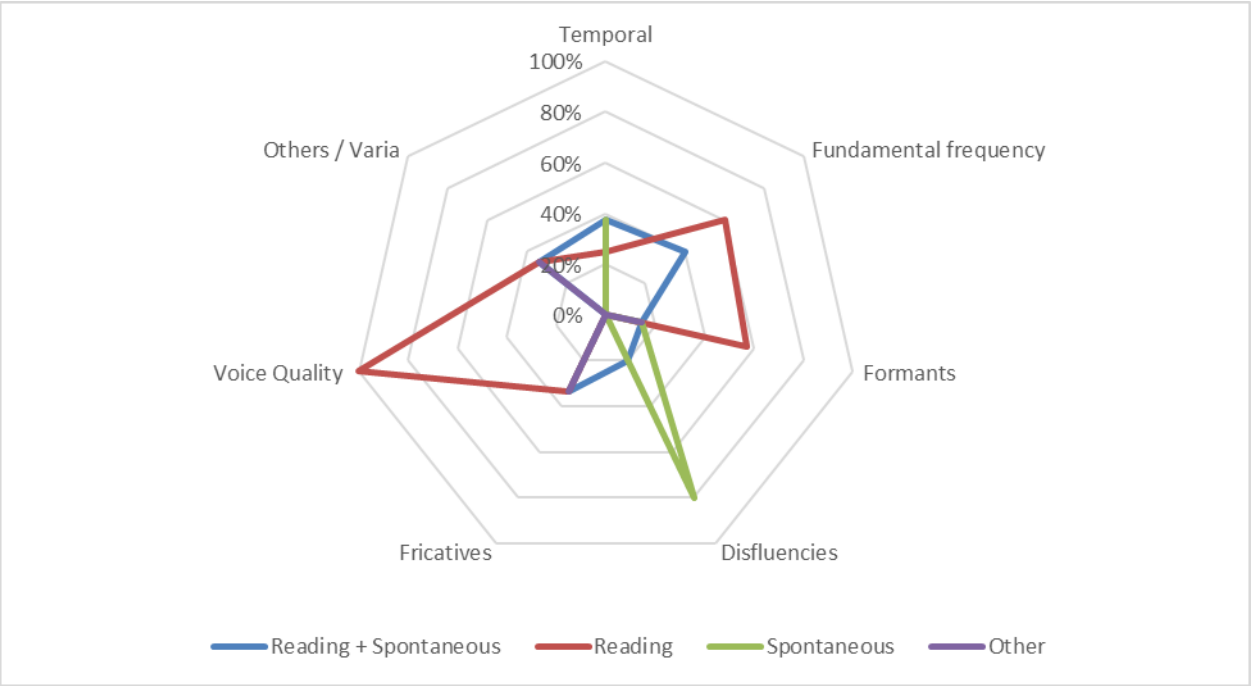


**Figure 2:** Distribution of speaking styles (reading and spontaneous, reading, spontaneous, and other) analyzed per parameter group: temporal parameters, fundamental frequency, formants, disfluencies, fricatives, voice quality and others/varia.

## 4.2. Intra-speaker variation factors

As noted above, the first factor inducing intra-speaker variation is speaking style. An important percentage of the analyzed studies (26.48 %) are concerned with how certain voice feature varies in a speaker from text reading to spontaneous speech. This is probably due to most common intra-speaker variation factor across studies, regardless of the voice parameter analyzed. Comparing how people vary in their production of certain sound or the occurrence of a particular speech phenomenon has been considered important in forensic phonetic studies to ascertain how robust that parameter is when comparing known and unknown voice recordings, when one consists of a spontaneous conversation (typically the offender recording) and the other one (typically the suspect recording) consists of read speech. This first type of intra-speaker variation was considered in three studies of the first group (temporal parameters: 28, 38, 39), two studies of the second group (fundamental frequency: 40, 41), and one study of the third group (formants: 33).

The studies of the first group point to the consistency of temporal-rhythmic features within speakers across speaking styles. Nonetheless, the rate of articulation in spontaneous speech is faster than that in read speech, and there is a correlation between the two rates: speakers with a faster speaking rate have also a faster reading rate (28). In the same way, the studies of the second group suggest that F0 features show a low degree of variability within speakers when considering the two distinct speaking styles. In contrast, the only study of the third group comparing formants in reading and spontaneous style shows disparate results: there is considerable intra-speaker variation across tasks when considering F3, while the values for F1 and F2 remained relatively consistent across tasks.

These results could have implications for deepfake detection, depending on how synthetic voices are created; basically, depending on the speaking style in which the pronunciation models for these artificial voices are based on.

As stated in the introduction, most forensic studies focus on comparing between-speaker and within-speaker variation, trying to find high between-speaker variation and low within-speaker variation, which is the desirable situation in FVC. Having reviewed the most commonly analyzed variation factor (speaking style), and finding that most voice parameters are robust taking it into account, we have extracted from Table 4 other factors potentially inducing intra-speaker variation (Table 5).

**Table 5:** Factors inducing intra-speaker variation

| *Factors* | *Type of acoustic parameters [parameter]* | *Paper reference* |
|---|---|---|
| Convergence with interlocutor | Temporal [speech rate] | 27 |
| Phrase length | Temporal [articulation rate] | 28 |
| Bilingualism | Temporal [speech rate] and F0 | 42 |
| Vocal effort | Fundamental frequency [F0varco] | 41 |
| Voice disguise | Fundamental frequency [F0 and F0SD] | 43 |
| Stressed (vs unstressed) syllable | Formants [F1, F2, and F3 of /aI/] | 31 |
| Content (vs function) word | Fundamental frequency and formants | 32 |
| Aging | Formants [F1] | 18 |
| Drunkenness | Disfluencies | 44 |

All of the factors listed above correspond to real factors of human variation that can be found in daily conversations, and that we deem hard to recreate in a realistic way in AI-generated voices.

Other sources of intra-speaker variation can occur, which do not depend on a specific factor but on the inherent variability of the sounds or parameters considered. For instance, Nadeu and Renwick (34) found that Catalan speakers are not always consistent in their realization of midvowels /e/ vs. /ɛ/; /o/vs./ɔ/). Also, Zhang, van de Weijer and Cui (17) compared the first four vowel formants (F1, F2, F3 y F4) of monophthongs, diphthongs and triphthongs in Chinese, without considering other potentially intra-speaker variation factors, and found that greater individual differences were found on F3 and F4.

## 5. Conclusion

Research on audio deepfake detection is key for tackling a number of societal challenges today, to name a few: distinguishing real news from fake content, or authenticating voice recordings in legal contexts, particularly when potentially fake recordings constitute forensic evidence. In order to answer the question of whether a voice is of human origin or created with AI techniques, experts need to know which voice characteristics to examine acoustically. To the best of our knowledge, the choice of such features is quite unguided today, since different experts examine disparate voice parameters, which typically amount to only a few ones, among all the possible dimensions in which audio can be examined, and those features are usually chosen on the basis that they are known to work well for distinguishing speakers; that is, human vs. speakers, but comparing human vs. deepfake is a different task.

Our systematic review departs from a clear hypothesis, which is that human voices present larger intra-speaker variation than deepfakes. Although that remains to be tested empirically, the first step, as well as the aim of this systematic review, has been examining the phonetic-acoustic literature that have delved into intra-speaker variation in any voice and speech aspect, and considering any possible language, with the purpose of summarizing and qualitatively analyzing

the published studies on the topic of intra-speaker variation in human voice. This has allowed us to shed some light on how humans vary in a range of acoustic features, which hopefully will distinguish us from artificially-generated voices.

Our results show that human voices are prone to intra-speaker variation for a large number of factors. Most investigations have focused on variation due to speaking style; namely, changing from reading a text passage to speaking spontaneously. Although some papers point to the consistency of voice parameters (mostly temporal and F0 features) across speaking styles, some other papers show that voice features can indeed change from reading a text to talking spontaneously. On the one hand, this has implications for deepfake detection if artificial voices are based on text-to-speech synthesis that rely strongly on reading pronunciation models. Human interaction is conversational and spontaneous. It does not consist of people reading. If the differences in such style-changing features are perceptually salient, one could distinguish deepfakes for not sounding 'spontaneous' enough. If such differences are not perceptually salient, at least the differences should be detectable by examining the speech signal acoustically. So, this first type of intra-speaker variation due to changes in speaking style could be important to know which voice parameters experts should focus on when trying to detect deepfakes.

On the other hand, we consider that the parameters that our systematic review has shown to be robust and useful to distinguish humans because they present low intra-speaker variation (particularly when it co-occurs with high inter-speaker variation) will not constitute necessarily useful voice features for distinguishing real voices from deepfakes, since lack of acoustic variation seem easily replicable for synthetic audio creation.

From the results of our systematic review, we can draw that —besides speaking style— there are, at least, eight possible factors inducing intra-speaker variation in real human voices. Our results show that these factors influence acoustic parameters of all possible sorts: temporal features, fundamental frequency, formants and disfluencies (Table 5).

Therefore, we propose that when analyzing a suspected deepfake from a traditional phonetic-acoustic perspective (i.e., without using an automatic system based on a black box model in which we do not know what its decisions are based on), instead of an unguided choice of phonetic features, experts could locate and acoustically analyze all the range of phonetic features that our systematic review has shown to exhibit large intra-speaker variation. Of course, some of the factors could be strongly context-dependent, like aging, drunkenness, voice disguise or bilingualism. However, most of the remaining factors are commonly found elements in real-life conversations (convergence with our different interlocutors and vocal effort) or they are linguistic conditions that could be examined (phrase length, differences between stressed and unstressed syllables, whether a particular sound occur in a content word or a function word, etc.).

While spoofing attacks using audio deepfakes typically occur with very short audios, trying to obtain longer audios from the person who potentially has stolen another person's voice identity (and trying to interact conversationally with this person/AI) would be the ideal situation to be able to obtain and measure most of the voice features that our systematic review has shown to present large intra-speaker variation.

All in all, we hope that deepfake detection systems will be able to nourish from the acoustic information derived from this systematic review. Indeed, we intend to apply the findings of this review in future experimental studies to advance state-of-the-art deepfake detection systems. In this way, systems will base their architectures on phonetic-informed parameters, in a so called bottom-up approach, and not only be based on deep neural networks.

Deepfake detection system nowadays work rather like a black box, with a large number of learnable neural layers, but scarce knowledge of what systems base their decisions on. Therefore, there is a constant call among experts to improve the interpretability of detection results (61), on

the one hand, and to improve automatic systems with expert knowledge, such as phonetic-acoustic expertise. Of course, deepfake data has to be collected, analyzed and confirmed that it has a smaller intra-speaker, which is our hypothesis. This hypothesis must be tested on deepfake corpora; meanwhile, phonetics-informed, expert-guided PAD remains underrepresented in current studies. As a case in point, a recent study by Yang et al. (58) confirms what we stated in the introduction and what have motivated our systematic review, when they state that "[*there is a*] need to approach audio deepfake detection using methods that are distinct from those employed in traditional FVC".

## 6. Acknowledgement

We thank the grant PID2021-124995OA-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

**REFERENCES**

1. Barrington, S., Cooper, E.A. & Farid, H. People are poorly equipped to detect AI-powered voice clones. *Scientific Reports 15*, 11004 (2025). https://doi.org/10.1038/s41598-025-94170-3

2. Ciancaglini V, Gibson C, Sancho D, McCarthy O, Eira M, Amann P, Klayn A, Malicious Uses and Abuses of Artificial Intelligence. Trend Micro Research, United Nations Interregional Crime and Justice Research Institute & Europol's European Cybercrime Centre; 2020 November 19.

3. Pfefferkorn R. Too Good to Be True? "Deepfakes" Pose a New Challenge for Trial Courts. NWLawyer, Washington State Bar Association; 2019 September.

4. McPeak A. The threat of deepfakes in litigation: Raising the authentication bar to combat falsehood. Vanderbilt Journal of Entertainment and Technology Law. 2020 23: 433–450.

5. Z. Khanjani, G. Watson, and V. P. Janeja: "Audio deepfakes: A survey," Frontiers in Big Data, vol. 5, pp.1-24, 2023.

6. Yamagishi, J., Todisco, M., Sahidullah, M., Delgado, H., Wang, X., Evans, N., ... & Nautsch, A. (2019). Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *ASV Spoof*, *13*.

7. WIRED. (2023, October 3). *Slovakia's election deepfakes show AI is a danger to democracy.* WIRED. https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy

8. Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A., & Böttinger, K. (2022). Does audio deepfake detection generalize?. *arXiv preprint arXiv:2203.16263*.

9. Jessen M. Forensic Voice Comparison. In Visconti J, editor. Handbook of Communication in the Legal Sphere. Berlín: De Gruyter Mouton, 2018. http://dx.doi.org/10.1515/9781614514664-012

10. French P, Stevens L. Forensic speech science. In Jones MJ and Knight RA, Editors. Bloomsbury companion to Phonetics. Londres: Continuum, 2005.

11. Jiménez-Peña J, Torres-Castillo FA, Cueva-Sánchez OE. Comparación forense de voces: un estudio preliminar sobre las diferencias entre una voz natural y una voz artificial para la investigación judicial. Revista Oficial del Poder Judicial. 2024 16 (21): 53–81. https://doi.org/10.35292/ropj.v16i21.881

12. Pellegrino E, Kathiresan T, Roswandovitz C, Fruholz S, Dellwo V. Can prosody be the key to spot fake voices? Acoustic and automatic speaker verification analyses on digital and natural voices. Paper presented at the 28th Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA). Istanbul; 20019 July.

13. Tan CB, Hijazi MHA, Khamis N, Zainol Z, Coenen F, Gani A. A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction. Multimedia Tools and Applications. 2021; 80 (21): 32725–32762.

14. Wolf J, Efficient acoustic parameters for speaker recognition. The Journal of the Acoustical Society of America. 1972; 51(6B): 2044–2056.

15. Nolan F. The phonetic bases of speaker recognition. Cambridge: Cambridge University Press; 1983. http://dx.doi.org/10.1016/0167-6393(87)90039-2

16. Kinnunen T, Li H. An overview of text-independent speaker recognition: from features to supervectors. Speech Communication. 2010; 52(1): 12–40.
17. Zhang C, van de Weijer J, Cui J. Intra- and inter-speaker variations of formant pattern for lateral syllables in Standard Chinese. Forensic science international. 2006; 158(2-3): 117–124. https://doi.org/10.1016/j.forsciint.2005.04.043
18. Rhodes R. Aging effects on voice features used in forensic speaker comparison. International Journal of Speech, Language and the Law. 2017; 24(2): 177–199. https://doi.org/10.1558/ijsll.34096
19. Ross A, Corley M, Lai C. Is there an uncanny valley for speech? Investigating listeners' evaluations of realistic TTS voices. Proc. Speech Prosody 2024. 2024; 1115-1119. https://doi.org/10.21437/SpeechProsody.2024-225
20. Bell A. Style in dialogue: Bakhtin and sociolinguistic theory. In Bayley R and Lucas C, editors. Sociolinguistic variation: theories, methods, and applications. Cambridge: Cambridge University Press; 2007, p. 90–109.
21. Bülow L, Pfenninger SE. Introduction: Reconciling approaches to intra-individual variation in psycholinguistics and variationist sociolinguistics. Linguistics Vanguard. 2021, 7(s2), 20200027.
22. Coupland N. Language, situation, and the relational self: Theorizing dialect-style in sociolinguistics. In Eckert P and Rickford JR, editors. Style and sociolinguistic variation. Cambridge: Cambridge University Press; 2001, p. 185–210.
23. Labov W. Sociolinguistic patterns. Philadelphia: University of Pennsylvania Press; 1972.
24. Covidence systematic review software. Veritas Health Innovation, Melbourne, Australia; [Accessed: February 2023]. Available at: www.covidence.org.
25. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol. 2008; 8: 45. https://doi.org/10.1186/1471-2288-8-45
26. Critical Appraisal Skills Programme: CASP Qualitative Checklist. [online]; 2018 [Accessed: 08/08/2024]. Available at: https://casp-uk.net/checklists/casp-qualitative-studies-checklist.pdf
27. Priva UC, Edelist L, Gleason E. Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. Journal of the Acoustical Society of America. 2017; 141(5): 2989-2996. https://doi.org/10.1121/1.4982199
28. Jacewicz E, Fox RA, Wei L. Between-speaker and within-speaker variation in speech tempo of American English. Journal of the Acoustical Society of America. 2010; 128(2): 839-850. https://doi.org/10.1121/1.3459842
29. Quene H. Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. Journal of the Acoustical Society of America. 2008; 123(2): 1104-1113. https://doi.org/10.1121/1.2821762
30. Eriksson EJ, Sullivan KPH. An investigation of the effectiveness of a Swedish glide + vowel segment for speaker discrimination. International Journal of Speech, Language and the Law. 2008; 15(1): 51-66. https://doi.org/10.1558/ijsll.v15i1.51
31. McDougall K. Speaker-specific formant dynamics: An experiment on Australian English /aI/. International Journal of Speech, Language and the Law. 2004; 11(1): 103-130. https://doi.org/10.1558/sll.2004.11.1.103
32. Heeren WFL. The effect of word class on speaker-dependent information in the Standard Dutch vowel/a:/. Journal of the Acoustical Society of America. 2020; 148(4): 2028-2039. https://doi.org/10.1121/10.0002173
33. Kuo C, Weismer G. Vowel reduction across tasks for male speakers of American English. Journal of the Acoustical Society of America. 2016; 140(1): 369-383. https://doi.org/10.1121/1.4955310
34. Nadeu M, Renwick M. Variation in the lexical distribution and implementation of phonetically similar phonemes in Catalan. Journal of Phonetics. 2016; 58: 22-47. https://doi.org/10.1016/j.wocn.2016.05.003
35. Harrington L, Rhodes R, Hughes, V. Style variability in disfluency analysis for forensic speaker comparison. International Journal of Speech, Language and the Law. 2021; 28(1): 31-58. https://doi.org/10.1558/ijsll.20214
36. Hughes V, Wood S, Foulkes P. Strength of forensic voice comparison evidence from the acoustics of filled pauses. International Journal of Speech, Language and the Law. 2016; 23(1): 99-132. https://doi.org/10.1558/ijsll.v23i1.29874
37. McDougall K, Duckworth M. Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of Standard Southern British English. International Journal of Speech, Language and the Law. 2018; 25(2): 205-230. https://doi.org/10.1558/ijsll.37241
38. Dellwo V, Leemann A, Kolly MJ. Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. Journal of the Acoustical Society of America. 2015; 137(3): 1513-1528. https://doi.org/10.1121/1.4906837

39. Leemann A, Kolly MJ, Dellwo V. Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. Forensic Science International. 2014; 238: 59-67. https://doi.org/10.1016/j.forsciint.2014.02.019
40. Leeman A, Mixdorff H, O'Reilly M, Kolly MJ, Dellwo V. Speaker-individuality in Fujisaki model f0 features: Implications for forensic voice comparison. International Journal of Speech, Language and the Law. 2015; 21(2): 343-370. https://doi.org/10.1558/ijsll.v21i2.343
41. Jessen M, Koster O, Gfroerer S. Influence of vocal effort on average and variability of fundamental frequency. International Journal of Speech, Language and the Law. 2005; 12(2): 174-213. https://doi.org/10.1558/sll.2005.12.2.174
42. Lee B, Sidtis D. The bilingual voice: Vocal characteristics when speaking two languages across speech tasks. Speech Language and Hearing. 2017; 20(3): 174-185. https://doi.org/10.1080/2050571X.2016.1273572
43. Künzel HJ. Effects of voice disguise on speaking fundamental frequency. International Journal of Speech, Language and the law. 2000; 7(2): 150-179.
44. Schiel F, Heinrich C. Disfluencies in the speech of intoxicated speakers. International Journal of Speech, Language and the Law. 2015; 22(1): 19-34. https://doi.org/10.1558/ijsll.v22i1.24767
45. Asadi H, Nourbakhsh M, He L, Pellegrino E, Dellwo V. Between-speaker rhythmic variability is not dependent on language rhythm, as evidence from Persia reveals. International Journal of Speech, Language and the Law. 2018; 25(2),: 151-174. https://doi.org/10.1558/ijsll.37110
46. Cavalcanti JC, Eriksson A, Barbosa PA. Multi-parametric analysis of speech timing in inter-talker identical twin pairs and cross-pair comparisons: Some forensic implications. Plos One. 2020; 17(1). https://doi.org/10.1371/journal.pone.0262800
47. He L, Dellwo V. The role of syllable intensity in between-speaker rhythmic variability. International Journal of Speech, Language and the Law. 2016; 23(2): 243-273. https://doi.org/10.1558/ijsll.v23i2.30345
48. Atkinson J. Interspeaker and intraspeaker variability in fundamental voice frequency. Journal of the Acoustical Society of America. 1976;60(2): 440-445. https://doi.org/10.1121/1.381101
49. Gandour J, Potisuk S, Ponglorpisit S, Dechongkit S. Interspeaker and intraspeaker variability in fundamental-frequency of thai tones. Speech Communication. 1991; 10(4); 355-372. https://doi.org/10.1016/0167-6393(91)90003-C
50. Correa J, Rodriguez L. (2018). Phonetic reduction of the consonant sequence /-st-/ in Bogota Spanish. Estudios Filologicos. 2018; 62: 193-214. https://doi.org/10.4067/S0071-17132018000200193
51. Haley K, Seelinger E, Mandulak K, Zajac D. Evaluating the spectral distinction between sibilant fricatives through a speaker-centered approach. Journal of Phonetics. 2010; 38(4): 548-554. https://doi.org/10.1016/j.wocn.2010.07.006
52. Munson B. A method for studying variability in fricatives using dynamic measures of spectral mean. Journal of the Acoustical Society of America. 2010; 110(2): 1203-1206. https://doi.org/10.1121/1.1387093
53. Harmegnies, B., & Landercy, A. (1988). Intra-speaker variability of the long term speech spectrum. *Speech communication*, 7(1), 81-86.
54. Jessen M. Speaker-specific information in voice quality parameters. International Journal of Speech, Language and the Law. 1997; 4(1): 84-103. https://doi.org/10.1558/ijsll.v4i1.84
55. Lee Y, Keating P, Kreiman J. Acoustic voice variation within and between speakers. Journal of the Acoustical Society of America. 2019; 146(3): 1568-1579. https://doi.org/10.1121/1.5125134
56. Lindsey G, Hirson A. Variable robustness of nonstandard /r/ in English: Evidence from accent disguise. International Journal of Speech, Language and the Law. 1999; 6(2): 278-289. https://doi.org/10.1558/sll.1999.6.2.278
57. Wassink A, Wright R, Franklin A. Intraspeaker variability in vowel production: An investigation of motherese, hyperspeech, and Lombard speech in Jamaican speakers. Journal of Phonetics. 2007; 35(3): 363-379. https://doi.org/10.1016/j.wocn.2006.07.002
58. Weirich M, Simpson A. Effects of Gender, Parental Role, and Time on Infant- and Adult-Directed Read and Spontaneous Speech. Journal of Speech, Language and Hearing Research. 2019; 62(11): 4001-4014. https://doi.org/10.1044/2019_JSLHR-S-19-0047
59. Amino K, Osanai T. Speaker characteristics that appear in vowel nasalisation and their change over time. Acoustical Science and Technology. 2012; 33(2): 96-105. https://doi.org/10.1250/ast.33.96
60. Galata V, Spreafico L, Vietti A, Kaland C. An acoustic analysis of /r/ in Tyrolean (WOS:000409394400208). 2016; 1002-1006. https://doi.org/10.21437/Interspeech.2016-434
61. Yi J, Wang C, Tao J, Zhang X, Zhang CY, Zhao Y. Audio deepfake detection: A survey. arXiv preprint arXiv. 2023; 2308.14970.

62. Yang T, Sun C, Lyu S, Rose P. Forensic deepfake audio detection using segmental speech features. arXiv preprint arXiv. 2025: 2505.13847. https://doi.org/10.48550/arXiv.2505.13847
63. San Segundo, E. (2024). Systematic review protocol. A qualitative systematic review of intra-speaker variation in the human voice. Zenodo. https://doi.org/10.5281/zenodo.13904591.