Evaluating Transcripts of Poor-Quality Forensic Audio: Sine-Wave Speech and Forensic Audio

Eleanor Kettle, Helen Fraser

Research Hub for Language in Forensic Evidence, University of Melbourne
eleanor.kettle@student.unimelb.edu.au; helen.fraser@unimelb.edu.au

Abstract

Sine-wave speech (SWS), created by combining sine-waves that track formants, initially sounds completely unlike speech, but priming listeners with the original sentence makes the words 'pop out' dramatically. Forensic audio shows a similar pop-out effect when a transcript is given – even if the transcript is inaccurate. In order to explore some of the nuances of the priming effect, this paper reports a new SWS experiment, in which the original sentence was only suggested, not given as the 'right answer'. Results are discussed in the context of forensic transcription, asking how responses can be evaluated in the absence of 'ground truth'.

Index Terms: sine-wave speech; forensic transcription; human speech perception

1. Introduction

Forensic speech recordings provide powerful evidence in criminal trials, but the audio is often of such poor quality that the jury needs a transcript to understand it. Current Australian law allows transcripts to be provided by police investigating the case [1]. To protect the jury from being misled, lawyers and judges check the transcript against the audio. This is ineffective, as listeners can be primed to hear in line with a transcript even if it is inaccurate [2] – see 90-second video at forensictranscription.net.au.

ASSTA is part of a 'call to action' asking the law to recognize forensic transcription as a science. This raises an interesting topic for speech science: how to reliably evaluate the accuracy of a transcript of poor-quality audio *in the absence of a known 'right answer', or ground truth*. The present paper explores some relevant issues by reviewing a famous 1981 experiment studying the perception of 'sine-wave speech' [3], then adding to its results with findings from a new experiment. Experiment files are at: https://www.mrc-cbu.cam.ac.uk/personal/matt.davis/sine-wave-speech/.

2. Sine-wave speech

2.1. What is sine-wave speech?

Sine-wave speech (SWS) was developed at Haskins Laboratories in the 1970s to provide non-speech files for use in experiments investigating human speech perception [4]. It is produced by combining two or three sine-waves that track the formants of an utterance. On first hearing, most people find it completely unlike speech: a series of beeps and whistles suggestive of some electronic source. However, when they are primed with the original sentence, most listeners find the words 'pop out' from the meaningless noise. This effect was first published in the original 1981 study by Remez et al [3].

2.2. The 1981 study

The 1981 study used a recording of the sentence: 'Where were you a year ago', to create seven SWS versions (audio not available): one with sine-waves tracking all three formants; three with two sine-waves tracking formants 1 and 2, 1 and 3, 2 and 3; and three with single sine-waves.

Each of these seven SWS versions were played under three conditions: Condition 1: participants given no information; Condition 2: participants told it was computer speech and asked if they could discern the content; Condition 3: participants told it was 'Where were you a year ago' produced by a computer, and asked how clear they found it.

One sentence, seven versions, and three conditions gives 21 experiments, deployed to 21 different groups of 18 undergraduate psychology students. Only versions with sine-waves tracking (a) F1-3, or (b) F1-2, gave useful results, so we focus on those here, for a total of 36 participants.

2.3. Results of the 1981 study

In Condition 1 (no information), only two of 36 participants, both listening to the 3-wave version, heard the full sentence. Nine (25%) thought it could be human or computer speech with various distortions. The others heard 'science fiction noise', 'radio interference' or other non-speech noises. In Condition 2 (speech suggested), nine (25%) heard the full sentence; 17 (47%) correctly heard varying numbers of syllables (average around 2.5 out of 7 total syllables); and ten (28%) still heard no words. In Condition 3 (sentence suggested), most listeners heard most words, with higher scores for the 3-wave version.

To the researchers, this showed that words could be heard in the absence of speech cues previously thought to be necessary to retrieve the vocal tract configuration that had produced the sounds. 'The results of this study cannot be explained within the framework of existing theories of speech perception, for the tones contained none of the elemental acoustic cues typically held to underlie speech perception'. (p.949)

2.4. Top-down vs bottom up

Since then, these results have generally been interpreted as strong support for the powerful role played by top-down processing in speech perception – though more sophisticated bottom-up processing, such as auditory scene analysis, has been defended by some [5].

A particularly thorough argument for the role of top-down processing was made in a 2007 paper by Davis and Johnsrude [6], reporting the 1981 results in the context of many other top-down effects: '… evidence supports highly interactive

processes with top-down information flow often driving and constraining interpretation.' (p.133)

A strong role for top-down processing is also supported by research on forensic transcription. However this work indicates that *inaccurate* transcripts can prime a 'pop out' effect as easily as *accurate* transcripts can. This led us to ask some questions about the 1981 results that have not yet, to our knowledge, been addressed in the SWS literature.

2.5. Some questions not yet addressed by SWS research

It is notable that SWS researchers have focused on the dramatic increase in participants who *do* hear the original sentence after it is suggested. This trend has been particularly prominent since 2007, when Davis and colleagues produced a set of SWS sentences which have received a great deal of media attention over many years [7-9]. SWS is typically presented via multimedia demonstrations, playing the incomprehensible SWS files, then the original sentence, then the SWS file again – which is now immediately accepted as, and assumed to be, the original sentence.

However this focus on an all-or-nothing effect is something of an overstatement. This is notable even in the 1981 experiment, where two participants heard the sentence before being told the file was speech; and after the suggestion, only 25% heard the sentence precisely as suggested, while 47% heard different words, and 28% heard no words at all.

The current experiment aimed to explore these nuances, by finding out what participants hear in SWS files if they are not provided with a single 'right answer'.

3. The current experiment

Our method was similar to that of the 1981 study, but also influenced by the design of forensic experiments such as [2].

We used the five 3-wave SWS files created by Davis and colleagues (link in Section 1). These were deployed via Qualtrics, under two sequential conditions. In Condition 1, participants were told they would be listening to 'distorted speech' and asked to write down what they heard (bypassing the 1981 condition of listening with no information). Condition 2 primed them with the suggestion (given via text, not audio) that 'Some people have suggested that they hear [the original sentence]'; and asked to listen again, as often as they wished, stating (a) whether they agreed; and (b) if not, what they heard instead.

A total of 52 participants completed the experiment. Most were undergraduates or recent graduates, with 69% aged 18-25; 25% aged 26-45 and 6% over 46. Most (31 of 52, 60%) were monolingual English speakers ('L1'), while 21 (40%) had various other language backgrounds ('other'), including 7 participants who self-reported as bilingual.

Participants were randomly divided into two groups (no statistically significant differences in age, gender, language background, headphone use, etc.). Group A listened first to all five SWS files under Condition 1, and then to each file under Condition 2. Group B listened to each file, first under Condition 1 and then under Condition 2. This gave them feedback on their perception of each file before they heard the next. All participants heard the files in the same order.

Participant responses were scored similarly to [10], with 3 = exactly like original; 2 = very close to original; 1 = some words but not close; 0 = no response. This gives 15 as the maximum possible score. Before-scores are the scores from Condition 1 (before original sentence suggested). After-scores are the scores from Condition 2 (after original sentence suggested – immediately after for Group B; at the end for Group A).

4. Results

4.1. Priming effect

Unsurprisingly, after-scores were markedly higher than before-scores (Figures 1 and 2), confirming that the suggestion had a strong priming effect, whether provided immediately or later. Notably, however, no suggestion was accepted by every participant. Indeed, some were explicitly rejected by more than half. Various factors appear to have affected the scores.

4.2. Participant effects

Before-scores varied greatly among participants (Figure 1). As in previous forensic experiments [10], this cannot be explained solely by language background. On average 'L1' before-scores were higher than 'other' ($p=0.049$), but some 'L1' participants did very poorly (e.g. two scored 0), while some 'other' participants did very well (1x12; 1x9; 3x7). Also as in previous studies (see also [11]), no other demographic characteristic we measured correlated with score. This suggests that the ability to decipher distorted speech with no contextual or textual suggestion reflects individual aptitude rather than specific demographic characteristics, in line with previous studies [10].

After-scores, though higher overall, were also variable (Figure 2), and also not correlated with language background or any other demographic characteristic we measured.

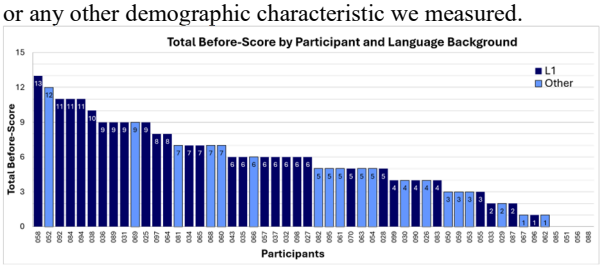

Figure 1: *Before-scores for all 52 participants.*

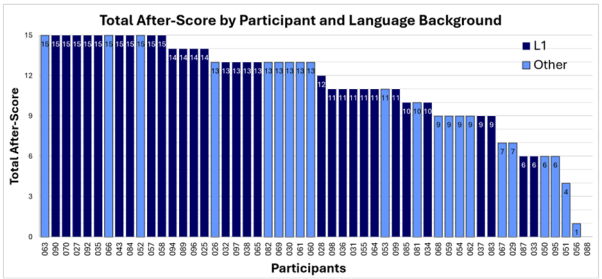

Figure 2: *After-scores for all 52 participants.*

4.3. Learning effect

Since 'other' here is a broad category with small numbers, and the effect of language background is not our present focus, all following results are reported only for the 31 'L1' participants (Group A: n=14; Group B: n=17). These smaller groups shared similar demographics to the overall participant field.

The 31 'L1' before-scores varied greatly by Group. As shown in Figure 3, for Group A (suggestion at end), highest scores were 11 and 9, most scored 6 or under, and two scored 0 (no words heard in any sentence). For Group B (suggestion

after each presentation), the highest score was 13, only six scored 6 or under, and none scored 0.

It seems Group B benefitted from the suggestion after each presentation, providing formative feedback that helped them understand subsequent SWS files before the suggestion. This SWS learning effect, which is very evident to those who experience multiple SWS examples, is also reported by others [6]. However, results are not uniform, even for Group B.

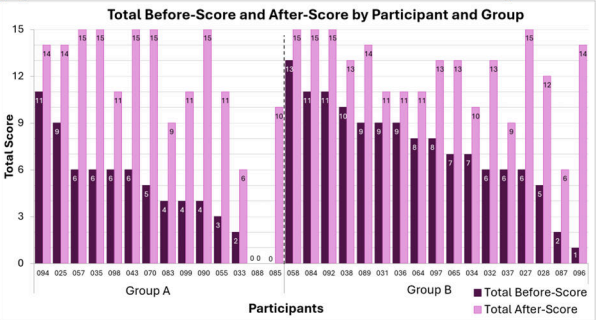

Figure 3: *Total before-score and after-score for each participant by group. Before-score shows a significant difference between groups (A mean = 4.7; B mean = 7.5; p=0.017). After-score shows no significant difference between groups (A mean = 11.5; B mean = 12.4; p=0.522).*

4.4. Between-sentence effects

Figures 4-6 show that a focus on average results masks considerable variation in responses to the original sentence uttered in each file:

1. It was a sunny day and the children were going to the park.
2. The camel was kept in a cage at the zoo.
3. The police returned to the museum.
4. The man read the newspaper at lunchtime.
5. He was sitting at his desk in his office.

Variation in before-scores is due in part to Group B's learning benefit creating higher scores for later files. Thus, while files 1 and 2 showed similar before-scores in both groups, scores for 4 and 5 were significantly higher for Group B. For Group A, by contrast, before-scores showed increasing scores of 0 for later files, suggesting these participants simply gave up trying.

However, both before- and after-scores suggest that files 1 and 5, perhaps also 4 to a lesser extent, are in some sense easier to understand than files 2 and 3. While files 1 and 5 were fully accepted by most (but not all) participants (87% and 84% respectively), only half fully accepted the suggested transcripts for Files 2 (52% overall) and 3 (48% overall).

At this stage it is not clear exactly what causes this differential effect. The first step for further exploration should be a follow-up study presenting the files in different orders.

4.5. Within-sentence effects

Further analysis revealed interesting differences in responses for individual key phrases within files. In this section, the number of correct responses is given as a percentage of all attempted responses (77% overall), setting aside those with 0 scores. The first observation (Figure 7) is that the end-phrase of each file was the most likely to be transcribed accurately (85% of attempted transcripts), while the accuracy for the start- (34%) and mid- phrases (37%) were notably lower. Thus 'the park' was accurately heard by 93% of those who attempted a transcript; 'the zoo' by 100%; 'the museum' by 86%; 'lunchtime' by 74%; and 'his office' by 63% (though this rises to 74% with the inclusion of 'the office' and 'office').

Again it is not entirely clear what causes this end-of-sentence advantage. It may simply be the end-focused intonation of these simple read sentences. Certainly more testing with a wider range of materials would be necessary before generalizing.

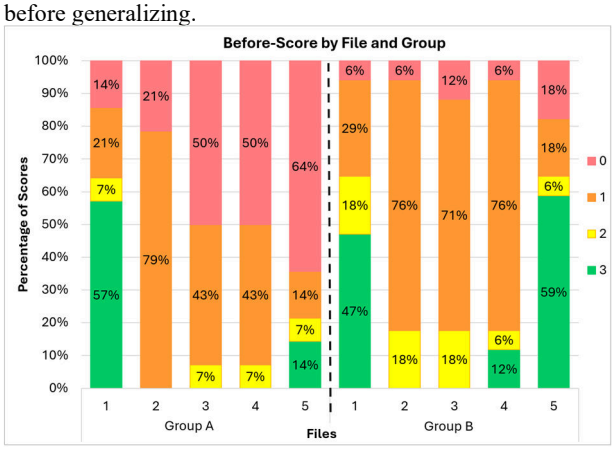

Figure 4: *Percentage of before-scores by file and group. Chi-square tests show that Groups A and B had similar scores for File 1 (p=0.669), File 2 (p=0.141), and File 3 (p=0.063); and significantly different scores for File 4 (p=0.031) and File 5 (p=0.04).*

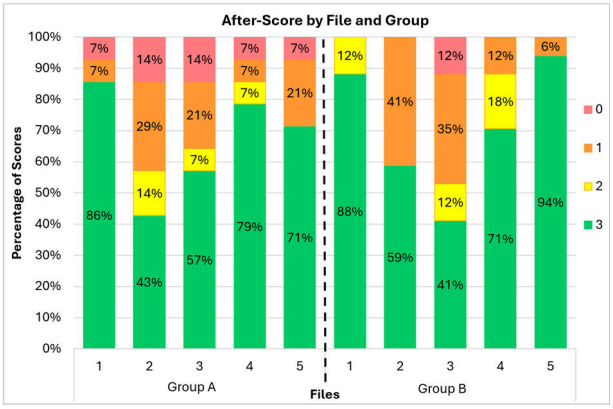

Figure 5: *Percentage of after-scores by File and Group. Chi-square tests by Group for each File show no significant difference (lowest p-value=0.1334).*

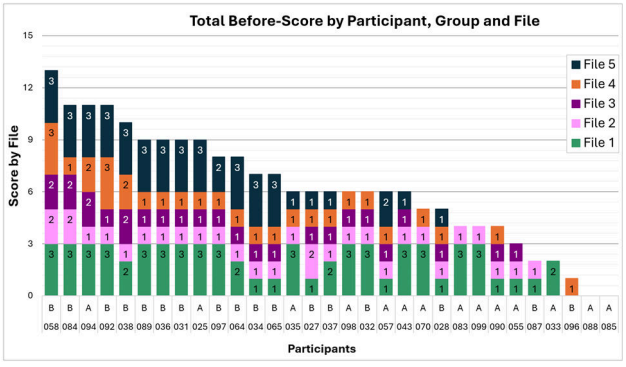

Figure 6: *Total before-scores for all 'L1' participants, shaded to show scores for each individual file.*

Particularly relevant in the present context is an analysis of common errors. Most notable is file 2, for which the correct transcription of the start-phrase 'The camel' was not given by any participants. Instead, three participants (11% of those attempting this file) gave 'the owl', 'the owls' or 'owl/arrow'; three gave 'now', two gave 'animal', and one gave 'cow'. In file 3, three (14%) gave 'please' as the start-phrase, while only two (9%) gave 'police'. For file 4, the mid-phrase 'the newspaper' was transcribed correctly by six (26%), but three gave 'museum', two gave 'music' and one gave 'muesli'.

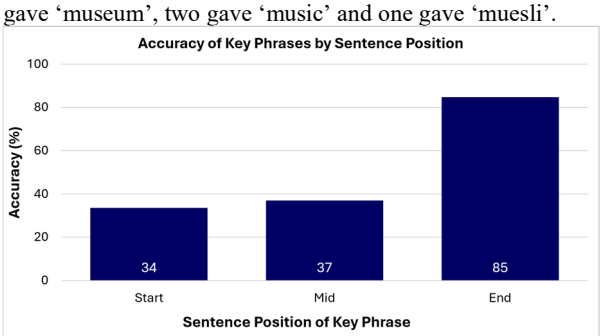

Figure 7: *Location of key phrases accurately transcribed before suggestion, as a percentage of attempted transcripts for each file.*

5. Discussion

Experiments with SWS and forensic transcription (FT) both yield observations of the strong power of a textual suggestion to make a meaningful utterance 'pop out' of unintelligible noise. However, while SWS research has focused on the *correct* utterance popping out, FT research has demonstrated that even a demonstrably inaccurate suggestion can create a confident pop-out perception. One of the difficult concepts speech scientists have to explain to the law is that just because listeners clearly hear particular words does not necessarily mean those words were the ones actually spoken [12]. As has long been known in phonetics, most stretches of speech are open to multiple interpretations even in clear speech (e.g. 'grey day' vs 'grade A') [13]; in poor-quality audio, interpretation depends heavily on listener expectations [14-15].

These considerations point to a major difference between SWS and FT research. All the results discussed above reflect scores predicated on 'accuracy' evaluated against a known 'right answer'. This 'ground truth' of course is not available in FT, prompting the question of how we could evaluate the responses in the current experiment in its absence.

One common suggestion is to use artificial intelligence [see 16]. Running the five current files through Whisper gave results as shown (scores in brackets), for a total before-score of 8, lower than the top third (30%) of the L1 participants:

1. *It was a sunny day and the children were going to the park.* (3)
2. *The owl is kept in a cage in the zoo.* (2)
3. *Please visit us at the museum.* (1)
4. *Allow me to meet you for lunchtime.* (1)
5. *He was sitting with his girlfriend, so I laughed.* (1)

Another common suggestion is to evaluate responses against the acoustics of the file. Here, however, while acoustic analysis can rule out some responses on the basis of rhythm or other features, SWS does not allow clear evaluation at a segmental level (see Figure 8). This is also a problem with forensic audio, where the quality can make evaluation via acoustic analysis difficult [17]. Experts are often more reliable in ruling out an incorrect suggestion than in giving a demonstrably correct interpretation.

This highlights another important difference between SWS and FT: the nature of the audio files. SWS starts from clear, read sentences, produced by a single speaker, then degraded via a regular process. FT originates from unmonitored conversation, and is degraded via multiple, variable and unknown factors. Only in rare cases could we expect a rapid learning effect such as that seen with feedback in the above results (and see also [11] for a learning effect without feedback).

Of course the controlled situations of SWS and similar experiments have many advantages, and have given speech science much useful knowledge. Forensic transcription, however, offers challenges that are also worthy of exploration by speech scientists. These factors make it valuable for researchers to experience and study the process of transcribing forensic-like audio under a range of forensic-like conditions [18], and especially to explore the conditions which are liable to make listeners, including analysts, confident but wrong.

Finally, it is worth noting that the challenge of understanding indistinct speech of unknown content in potentially misleading contexts is exactly the challenge that listeners face in real-life situations every day. Researching FT therefore potentially has implications for theoretical accounts of human speech perception.

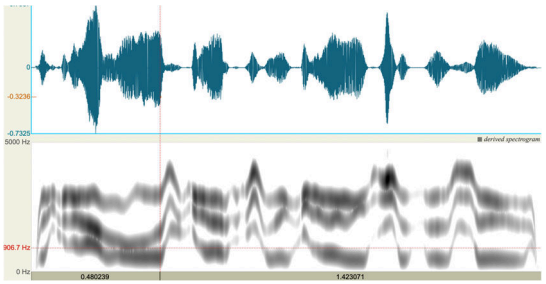

Figure 8: *Spectrogram of File 2, with location of 'The camel' indicated. While this is consistent with 'the camel', it could also support other interpretations.*

6. Conclusion

Experiments on sine-wave speech, like many other kinds of speech perception experiments, are typically carried out by researchers who know the true content of the audio, and are able to present systematically distorted samples to participants and systematically deprive them of relevant information. Such research can of course provide a great deal of useful knowledge about human speech perception. However it does have some risk of researchers getting 'locked in' to the 'correct response', without considering the validity of other responses, given the potential for the distorted acoustic signal to be interpreted in multiple ways.

The next step for this experiment is to continue investigating possible interpretations of sine-wave speech, with a particular focus on the effects of different feedback on the listeners' interpretations (such as presenting the original sentence; a plausible but misleading sentence; or both of these options).

7. References

[1] Fraser, H. “The development of legal procedures for using a transcript to assist the jury in understanding indistinct covert recordings used as evidence in Australian criminal trials: A history in three key cases”, Language and Law=Linguagem e Direito, 8(1), 59–75, 2021. doi: 0.21747/21833745/lanlaw/8_1a4

[2] Fraser, H. and Kinoshita, Y. “Injustice arising from the unnoticed power of priming: How lawyers and even judges can be misled by unreliable transcripts of indistinct forensic audio”, Criminal Law Journal, 45(3), 142–152, 2021. http://hdl.handle.net/11343/285048

[3] Remez R., Rubin, P., Pisoni, D. and Carrell, T., “Speech perception without traditional speech cues”, Science, 212(4497):947-950, 1981. http://www.jstor.org/stable/1685714

[4] Rubin, P. “SWS: an overview and history”. Unpublished Manuscript, 2005. https://static1.squarespace.com/static/6463dc0cc0fa823cfe25e5fd/t/648bc8f2464bdb011edbfaeb/1686882549081/SWSOverview2005.pdf

[5] Barker, J. and Cooke, M. “Is the sine-wave speech cocktail party worth attending?”, Speech Communication, 27:159-174, 1999, doi: 10.1016/S0167-6393(98)00081-8

[6] Davis, M. and Johnsrude, I. “Hearing speech sounds: Top-down influences on the interface between audition and speech perception”, Hearing Research, 229(1–2):132–147, 2007, doi: 10.1016/j.heares.2007.01.014

[7] Lawton, G. “Mind tricks: Ways to explore your brain”, New Scientist. 195(2622):34-41, 2007.

[8] Smith, C. and Critchlow, H. “Ketamine and schizophrenia”, The Naked Scientists podcast, 2013. https://www.thenakedscientists.com/articles/interviews/ketamine-and-schizophrenia

[9] Seth, A. “Your brain hallucinates your conscious reality” TED talk. 2017. https://www.ted.com/talks/anil_seth_your_brain_hallucinates_your_conscious_reality/

[10] Fraser, H., Loakes, D., Knoch, U., and Harrington, L. “Towards accountable evidence-based methods for producing reliable transcripts of indistinct forensic audio”, International Association of Forensic Phonetics and Acoustics (IAFPA). Zurich. 2023.

[11] Cooke, M., Scharenborg, O., and Meyer, B. T. “The time course of adaptation to distorted speech”, The Journal of the Acoustical Society of America, 151(4), 2636–2646, 2022, doi: 10.1121/10.0010235

[12] Fraser, H. “‘Assisting’ listeners to hear words that aren’t there: dangers in using police transcripts of indistinct covert recordings”, Australian Journal of Forensic Sciences, 2018, doi: 10.1080/00450618.2017.1340522

[13] Lehiste, I. Readings in Acoustic Phonetics. MIT Press, 1967.

[14] Repp, B. “Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception”, Psychological Bulletin 92(1) 81-110, 1982, doi: 10.1037/0033-2909.92.1.81

[15] Shockey, L., and Bond, Z. “What slips of the ear reveal about speech perception”. Linguistica Lettica, 22, 107–113, 2014.

[16] Loakes, D. “Automatic Speech Recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare?”, Frontiers in Communication, 9(Capturing Talk), 2024, doi: 10.3389/fcomm.2024.1281407

[17] French, P., and Fraser, H. “Why ‘ad hoc experts’ should not provide transcripts of indistinct forensic audio, and a proposal for a better approach”, Criminal Law Journal, 42(5), 298–302, 2018.

[18] Kettle, E. “Feel sorry for your ears”: Exploring challenges in transcribing speech with unknown content in unfamiliar varieties. Australian Linguistics Society conference, 2024.