**AES76-xxxx**

**DRAFT**

# AES standard for audio forensics - Speech Collection Guidelines for Speaker Recognition: Interviewing at a Temporary Location

Published by
**Audio Engineering Society, Inc.**


**Abstract**
This document specifies recommended practices for recording audio intended for use in forensic speaker recognition analyses, focusing on doing so at a temporary, non-laboratory location by possibly a non-professional in the forensic sciences. It includes recommendations for the physical preparation of the location, selection of appropriate recording hardware and audio formats, and possible methods for interviewers to elicit the desired type and amount of speech from subjects. It does not cover the methods used to analyze the resulting recordings and does not deal with details related to the handling, transmission, storage, or preservation of the collected data but will include a checklist to aid in the process.

**AES76-xxxx**

**DRAFT**

# AES standard for audio forensics - Speech Collection Guidelines for Speaker Recognition: Interviewing at a Temporary Location

Published by
**Audio Engineering Society, Inc.**


**Abstract**
This document specifies recommended practices for recording audio intended for use in forensic speaker recognition analyses, focusing on doing so at a temporary, non-laboratory location by possibly a non-professional in the forensic sciences. It includes recommendations for the physical preparation of the location, selection of appropriate recording hardware and audio formats, and possible methods for interviewers to elicit the desired type and amount of speech from subjects. It does not cover the methods used to analyze the resulting recordings and does not deal with details related to the handling, transmission, storage, or preservation of the collected data but will include a checklist to aid in the process.

## Contents

## Foreword

This foreword is not part of this document, AES-x253-20200708 AES standard for audio forensics -Speech Collection Guideline for Speaker Recognition: Audio Collection at a Temporary Location.

The document was originally produced by the Organization of Scientific Area Committees for Forensic Science, Digital/Multimedia Scientific Area Committee, Speaker Recognition Subcommittee (OSAC-SPEAKER), as a continuation from SWG-SPEAKER (Scientific Working Group – Speaker Recognition).

Original Authors of the document drafted and reviewed/approved first from SWG-SPEAKER and subsequently from USA OSAC-SPEAKER are: Stephen Gibbs (principal author), Darren Haddad (key contributing author)
Additional OSAC-SPEAKER Contributors: Walt Andrews, Jos Bouten, Joseph Campbell, Christopher Cieri, Kathleen Egan, Kevin Farrell, Patrick Gibbs, John Hansen, Aaron Lawson, David Marks, Johanna Morley, Geoffrey Morrison, Hirotaka Nakasone, Douglas Reynolds, Reva Schwartz, James Wayman. Other unnamed members of the OSAC/QIC committee.

This document was developed under project AES-X253, in the SC-03-12 task group, first led by John Hansen then by Marisa T. Déry.

Contributors in the task group include M.T. Déry, J. Hansen, R. Cabot, D. Boss, G. Reid, C. Lamschtein, K. McElveen, J. Owen, C. Grigoras, R. Maher, J. Woodgate, D. Bergfeld, K. Marr, S. Winner, D. Walsh, E.B. Brixen, S. Beck.

At the time of publication the AES SC-03-12 consisted of: E. Achren, A. Alexander, R. Alexanderson, S. Allen, W. Andrews, S. Bamford, A. Bartle, S. Beck, D. Bell, D. Bergfeld, T. Birkeland, M. Bittle, D. Boss, E. Brixen, N. Broner, J. Broyles, B. Brustad, W. Bulla, R. Cabot, S. Cain, A. Castner, J. Charles, P. Connolly, M. de Cunto, A. de Oliveira, M. Déry, W. Deutsch, T. Edwards, E. Enzinger, T. Erbe, J. Freytag, R. Gaskell, S. Gibbs, G. Gibson, R. Gorsline, J. Griffin, C. Grigoras, B. Gunel, H. Hacihabiboglu, D. Hallimore, J. Hansen, P. Harrison, S. Herla, C. Hordyk, G. Houle, B. Hunt, S. Jacobson, P. Kaeseler, F. Kelly, K. Kim, M. Kleiner, B. Koenig, W. Krafft, C. Lamschtein, T. Lewis, R. Maher, H. Malik, D. Marks, K. Marr, R. Max, K. McElveen, P. Merrill, G. Morrison, J. Narus, W. Newman, D. Notowitz, D. Nyberg, N. O'Brien, P. Olsson, J. Owen, T. Owen, P. Pegas, A. Pellicano, C. Peltier, M. Piper, F. Poza, E. Primeau, J. Ramirez, J. Reames, A. Reich, G. Reid, W. Reynolds, R. Rose, M. Roy, M. Sanfilipo, R. Schlesinger, B. Skye, C. Smith, J. Smith, L. Smrkovski, A. Stanley, J. Stathakis, R. Streicher, G. Stutchman, L. Tjellesen, R. Todd, J. Trš, J. Trautner, D. Tunstall, A. Tuzman, C. Van Winkle, Z. Varga, J. Vermeulen, J. Voetmann, G. Wallach, S. Walsh, J. Wheeler, L. Williams, S. Winner, J. Woodgate, D. Wujcik, J. Zjalic.

Marisa T. Déry
Chair, working group SC-03-12
2022-01-12

### Note on normative language

In AES standards documents, sentences containing the word “shall” are requirements for compliance with the document. Sentences containing the verb “should” are strong suggestions (recommendations). Sentences giving permission use the verb “may”. Sentences expressing a possibility use the verb “can”.

# DRAFT

# AES standard for audio forensics - Speech Collection Guidelines for Speaker Recognition: Interviewing at a Temporary Location

## 0 Introduction

### 0.1 General

In this document, the person overseeing the collection session will be designated the “interviewer” and the individual being recorded the “subject”. The intended audience for these guidelines is those interviewers called on to perform speech collection using portable equipment at a temporary location not originally designed or intended for audio recording.

The goal of speech collection is to collect an audio recording containing a combination of subject identifying information (such as their name, date of birth, etc.) and speech which can be used for future comparison against the speech of unknown speakers using unspecified speaker recognition methods. Although the specific method of speaker recognition is left undefined, automated/semi-automated computer-based methods were the primary driver for some of the specific parameters found in these guidelines.

These guidelines should be viewed as providing minimum requirements for usable speech collection in an operational or field environment and are not intended as data collection guidelines for research applications, speech intended for transcription, or other applications.

It is important that before implementation of these guidelines the user coordinate their activities with any elements of their parent organization which will be storing and using the collected data. Issues related to audio channel (microphone, telephone, radio, etc.), desired languages and dialects, data formats, etc. should be worked out beforehand to ensure that the data is maximally useful. Similarly, the proper storage and protection of personal identifying information (if collected) should also be coordinated as required within your organization prior to data collection.

### 0.2 Documentation conventions

Numerical values are decimal unless otherwise stated.

Where new terminology is first introduced in body text, the term will be set in an *italic typeface*.

# 1 Scope

These guidelines are intended to be one of a series and covers only one scenario - the collection of speech samples for speaker recognition at a temporary, non-laboratory location. One example of this would be the collection of speech samples from subjects during some type of field activity. The field activity could be associated with a range of purposes, such as law enforcement, intelligence, military or sociological. These guidelines presume portable resources and somewhat limited time to perform the collection. It also assumes that the interviewer is fluent in the subject's language, or that an interpreter is present who is both fluent in the interviewer's and subject's languages and cognizant of the goals of the interview.

These guidelines do not deal with details related to the handling, transmission, storage, or preservation of collected data. Specifically:

- It does not deal with any issues related to collection, storage, or protection of personal information.
- It does not recommend how to protect speaker recognition analysts from seeing or hearing subject identifying information which could result in biased analysis results.

It is the responsibility of the guideline's user to determine what their organization's rules and policies are on these matters and to tailor their implementation of these guidelines to comply with those rules and policies. It is also the guideline user's responsibility to learn how to operate the selected equipment, especially the placement orientation and distance of the microphone and the interpretation of any meters on the recording device.

These guidelines also do not deal with possibly important concerns such as personnel and equipment security, power sources for equipment, etc. The alleviation of such highly situation specific concerns is the responsibility of the guideline's user or their agency.

# 2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

**AESxx-1** *AESxx Part 1: AES standard for - PLACEHOLDER*. Audio Engineering Society, New York, NY., US.

# 3 Terms, definitions, and abbreviations

For the purposes of this document, the following terms, definitions, and abbreviations apply.

### 3.1 A/D

Analog-to-digital

### 3.2 AGC

Automatic Gain Control. A closed-loop regulating circuit which provides a controlled signal amplitude at its output, despite variation of the amplitude in the input signal.

### 3.3 Codec

Algorithm designed to encode or decode a stream of digital audio data.

### 3.4 Hz

Hertz is the international unit of measurement of frequency. One hertz corresponds to one cycle per second.

### 3.5 Mbyte

Megabyte (of digital storage) – this normally refers to 1024*1024 (1,048,576) bytes in digital applications though others may intend 1000*1000 or 1,000,000 bytes.

### 3.6 MD5

A widely used cryptographic hash function producing a 16-byte hash value, typically expressed in text format as a 32-digit hexadecimal number. MD5 is commonly used to verify data integrity.

### 3.7 min

Minute (of time)

### 3.8 MP3

A widely used audio file format which uses a lossy audio encoding algorithm defined in the MPEG-1 standard, Audio layer 3. The details of this standard are which is published as ISO/IEC 11172-3.

### 3.9 PCM

Pulse Code Modulation. Refers to a method of representing an analog audio waveform with a series of quantized digital sample values.

### 3.10 PCM-WAV

A version of the WAV file format which saves the data as uncompressed linear PCM samples with a standard RIFF header.

### 3.11 RIFF

The Resource Interchange File Format is a generic file container format which can be used to store audio data.

### 3.12 USB

Universal Serial Bus. Refers to a family of standardized computer peripheral interfaces.

### 3.13 WAV

A specific implementation of the RIFF file format for audio data.

### 3.14 WMA

Windows Media Audio. This refers to both a Microsoft proprietary audio file format and the audio codecs it uses. WMA can use either lossy or lossless codecs.

## 4 Speech Collection Scenario: Audio Collection at a Temporary Location

### 4.1 Collection Environment

The collection environment should:

- Be an indoor space as free as possible from background noises such as air conditioners, generators, fans, or other motorized or electrical devices. Avoid locations that have music, white noise, or other audio playing

in the background at any audio level. It may be necessary to turn the interference source off to fully mitigate it.

- Be a location that is not near outside traffic noise (human, animal, vehicular, or aircraft).

- Have a minimum of large, flat, hard sound-reflective surfaces which can cause reverberation and echoes. The effects of a reverberant room can be mitigated by hanging fabric (curtains, blankets, etc.) or other sound deadening materials on the walls or as dividers in the room.

- Allow the subject to be as comfortable as possible, preferably sitting, to lower cognitive/voice stress levels and to facilitate natural conversation.

### 4.2 Collection Equipment

Although high-quality digital audio recording equipment is preferred, recordings made with equipment meeting the minimum requirements detailed below should be useable by most speaker recognition methods. If there are multiple recording sites, the same type of recording system should be used at all collection locations to minimize recording differences due to equipment variation. Nothing in these requirements precludes the concurrent use of multiple recording devices if that is required for the intended application. An example of such a requirement would be to create a pair of recordings in which one is a high-quality reference while the other is condition-matched to a specific use case.

Recording devices should fulfill the following requirements:

- The speech should be recorded digitally and saved as uncompressed PCM data with at least 16-bit samples at a minimum rate of 16 kHz. The audio can be mono or stereo. Recording at a higher sample rate and bit depth is greatly preferred if that is possible. Many current devices support the recording of 16 or 24-bit samples at rates up to 48 kHz and storage of the recorded data in PCM-WAV format[1].

- For recordings made using laptop or other computers, it is preferred to use an external USB condenser microphone with an on-board analog-to-digital (A/D) converter. This is because the internal microphone or external microphones plugged into a “mic” port can pick up noise from internal circuitry.

- The subject’s microphone should ideally be a headset mic since:

    - it fixes the location of the mic with respect to the mouth.

    - it reduces interference from the interviewer’s speech and any background sounds.

    - the speaker will quickly forget about its presence.

  Otherwise, a microphone on a stable stand or tripod which places it at an appropriate distance from the subject for the type of microphone being used is acceptable. If the recording device/microphone is directional, it should be situated to best pick up the subject’s speech. If not chosen to mimic an operational scenario, the microphone should ideally provide a flat frequency response.

1 If the recording equipment used samples at rates lower than these recommendations, the recordings may still be suitable for some applications but must be seen as information-losing and thus suboptimal.

- A foam or fabric breath guard (commercial or improvised) should be used with the microphone to mitigate puffing noises from the subject's breath or burst type speech sounds such as stop consonants (i.e., /t/, /p/, /k/) if possible.

- The interviewer should have some indicator available on the recording device that shows that the audio is being recorded at an appropriate amplitude level and not too low (resulting in a noisy recording due to quantization effects) or too high (which causes clipping and thereby introduces nonlinear distortion into the audio stream). The relationship of appropriate sound levels and equipment indicator displays is not standard, therefore the indicator level providing acceptable audio should be determined by testing before the collection activity.

- There should be some method available to back-up the collected data, such as writing it to optical media, external hard drives, or USB thumb drives.

### 4.3 Speech Collection

After the collection environment and equipment have been arranged, the interviewer should record and audibly review an initial sample of test speech in the same recording environment and using the same equipment as for the collection to confirm that the equipment is working properly, and the audio quality meets the parameters discussed above. This may also expose other sources of noise not originally noted, such as the buzz of fluorescent lights or sounds from air handlers, which can be addressed as discussed in section 5.1. Once the setup is verified, it should be documented, ideally including the model identification and serial numbers of the equipment and a diagram or photographs of how it was connected and arranged.

Once recording begins, either the interviewer or the subject must provide a preamble with some subject identifying information along with the date, time and location of the recording session. The preamble information may be elicited with a set of fixed questions to provide identity information (such as name, address, and date of birth), or it may consist of the statement of a unique identifier, such as an identifying number which relates the subject to separately documented identifying information. If of value to the speaker recognition method used by the interviewer's organization, the identifying questions or statements could be tailored to provide specific known-text or phonetic content.

The type of speech (conversational, reading, preaching, etc.) recorded during the collection should ideally match what is expected to be in the unknown speech samples to be compared against in the future. This is also true of the language used. If the interviewer knows what type of speech and the language the subject recordings will be compared against, the collection should be designed to elicit those. These may not be possible to know beforehand, and the capture of conversational speech is discussed below as a general example.

During the recording, the interviewer should strive to elicit periods of conversational speech from the subject without interrupting or speaking at the same time as the subject.

Conversational speech could be elicited in multiple ways, such as:

- Engage the Speaker by discussing a subject of interest

- Asking open-ended questions or prompts. A list of possible questions is given in Appendix 1. This list is not exhaustive, and the interviewer should tailor any questions to be appropriate for the circumstances, the subject's culture, etc.

- Asking the subject to describe or interpret an image. These could be simple drawings of an object or scene, photographs of a general nature, or other images that have content understandable by the subject and which will elicit a conversational response. Some examples are provided in Appendix 2. Giving the subject a choice of several images gives them the freedom to choose one for themselves. The interviewer could ask

multiple questions about an image to elicit additional speech, such as “are there any dangerous things in the image?” Alternatively, the interviewer could circle some items in the image, and ask the subject to describe them.

- Ask the subject to discuss an article from a local newspaper, news website, or social media outlet.

Note – in the last two methods, the use of paper copies of the image, drawings or newspaper articles should be avoided since their movement can add undesired noises to the recording.

It can be expected that the longer the subject speaks conversationally (presuming that fatigue does not occur), the greater chance that they will become comfortable with the collection situation, resulting in a more “natural” speech sample. In case the subject does not engage in conversation, it is recommended that 20 or more additional questions similar to those in Appendix 1 be developed beforehand so that an adequate amount of speech can be collected.

The interviewer should avoid interjections while the subject is speaking (e.g., nodding to acknowledge the subject instead of saying "uh-huh").

The subject’s portion of the recording (including the identification segment, any answers to questions, and conversational speech) should contain a minimum of two minutes of speech and preferably up to five minutes. This is to be measured after the removal of speech from the interviewer, any noisy segments, and extended pauses.

After the completion of the recording session, the interviewer should document any comments on the collection (via handwritten notes, verbal recordings, etc.) prior to beginning the next session.

## Annex A (Informative) - Possible Known-Text Phrases for the Subject

Questions based on the following phrases can be used to capture subject-identifying information and elicit known text responses from the subject. They should also help encourage discussion which will fulfill the desire for the capture of conversational speech. If additional conversational speech is needed, questions should be designed to allow open-ended responses instead of short answers.

- “My name is [Subject name]”
- “I was born on [Subject date of birth]”
- “I was born in [Subject place of birth]” – this can include town, city, region, country, etc.
- “I currently live in [Current residence location] at [address]”
- “My current job/occupation is [Occupation]”
- “I am [married/single] …”
- “My [height/weight/eye color] is …
- “…my wife (husband) is named [name] …”
- “…and we have [#] children – who are [Names, ages, etc].”
- “My favorite pet’s name is [Boots, Garfield, etc]”
- “My favorite sport is [sport, ex: rugby, running, swimming, etc.]”
- “My first vehicle was a [Vehicle with details, ex: red 1970 Ford Mustang]”
- “The best phone number you can reach me at is [phone number]”

## Annex B (Informative) - Examples of Images to Elicit Speech Samples

The images below are examples of ones that could be used to elicit speech from subjects. As noted above, the images must be understandable by the subject and personally/socially acceptable to them.

(The four photographs below were acquired from the Wikimedia Commons and are attributed to their authors.)

*Fig.1 (*Geneva Rugby Cup - 20140808 - SF vs LOU. By Clement Bucco-Lechat.)

*Fig.2 (*A desert: The rain shadow region of Tirunelveli, India. By Arun Ganesh.)


*Fig.3* (Patagonia 4x4 Off Road Expeditions by Mil Outdoor Adventure, El Calafate, Provincia de Santa Cruz, Patagonia Argentina. By total13.net.)

*Fig.4 (Philippine rice terraces, Batad village. By Adi.simionov,)*

*Fig.5* (Camping/Outdoors. Provided by Jos Bouten.)

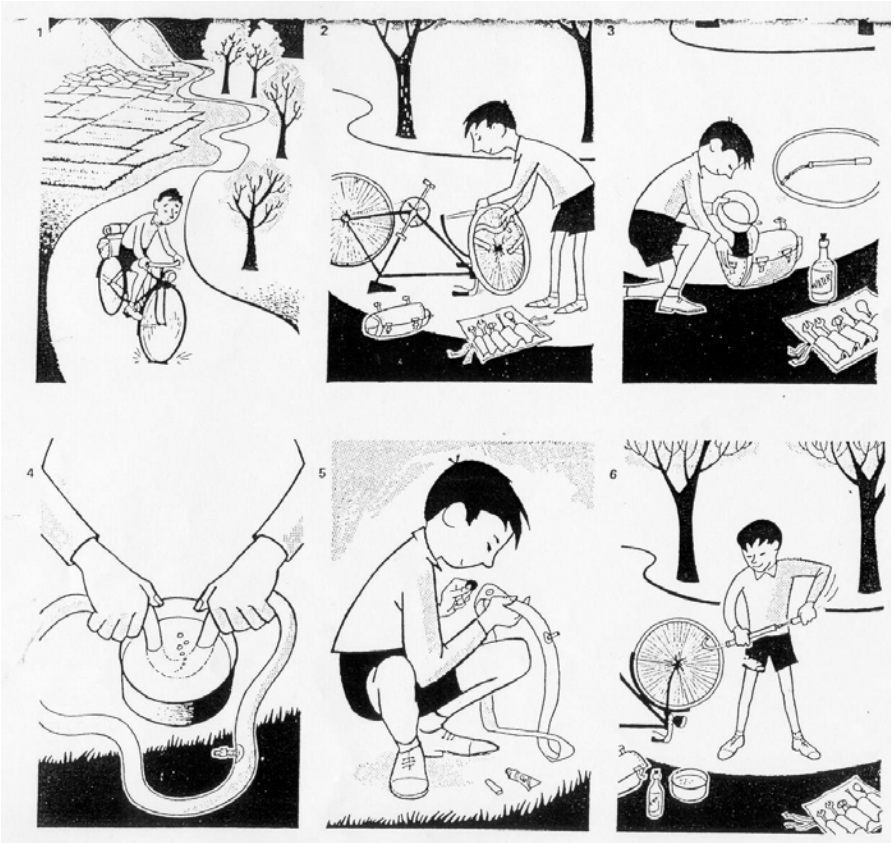



*Fig.6* (Fixing a bicycle tire. Provided by Jos Bouten.)

## Annex C (Informative) - Example of a Checklist

Relevant Case Data:

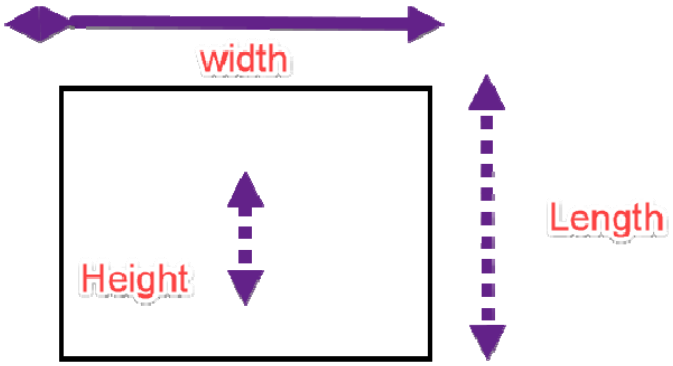
Date and Time:

Interviewer:

Speaker:

Location:

Room (if used):




Recording Device:

Microphone used (if external):

Microphone distance to the Speaker:

Physical State/ Health (e.g., Dehydration, Sleep Deprived):

Speaker Traits (e.g., State of Teeth, Stutter, Chemically Induced State):

Emotional State:

General images used to encourage spontaneous speech (if used):