Speaker Detection in a Forensic Environment: Recognizing the Limitations, Improving the Science

Alvin F Martin
alvin.martin@nist.gov

International Symposium on Forensic Science Error Management

Detection, Measurement and Mitigation

Arlington, VA

July 20-24, 2015

Outline

- Some forensic speaker detection history
- Crime scene challenges
- Factors affecting audio evidence quality
- Evaluation and its limitations
- NIST HASR tests
- Role of OSAC
- NIST Vendor SRE Evaluation

Some History

- Voiceprint identification – a term for speech spectrograms going back to the 60's - has often been taken to suggest that voice recognition is as reliable as that by fingerprint – NOT SO!
- Forensic detection may include
  - Spectrograms
  - Phonetics (preceded spectrograms, used in Lindbergh case)
  - Automatic speaker recognition (considerable recent advances)
- Warning presented at Eurospeech 2003
  - *at the present time, there is no scientific process that enables one to uniquely characterize a person's voice or to identify with absolute certainty an individual from his or her voice*
  - Joe Campbell, Doug Reynolds, et al., "Person Authentication by Voice: A Need for Caution"

Some History (cont'd)

- Changing judicial acceptance standards for scientific evidence
  - – Voiceprint identification was subject of many early cases – rulings both ways
  - – General standard has gone from Frye to Daubert (federal and many states)
    - Daubert criteria for theories and techniques: testability - peer-review & publication - known or potential error rate - general acceptance in relevant scientific community

Crime Scene Challenges

- Crime scene recordings are highly variable, often problematic, and different in major ways from known suspect recordings
  - Involve many types of emotion and stress
  - May involve substance abuse, physical or emotional violence, shouts and yelling
  - Speech utterances likely to be short
  - May be attempts at voice disguise
  - Background conditions uncontrolled, often highly noisy or reverberent
  - Recording equipment may be of poor or unknown quality or provenance

Crime Scene Challenges (cont'd)

- Practitioners should recognize when investigative results are possible and when not
  - Results may be significant or inconclusive
  - May provide investigative assistance or basis for court testimony

- Knowing when to "punt" – some likely reasons
  - Limited duration
  - High noise
  - Shouting/yelling/whispering
  - Non-speech/non-intelligibilty

Types of Factors Affecting Performance

- Extrinsic – external to the speaker
  - Channel – telephone, microphone, etc.
    - Multiple types, different handsets, cell, voip, etc.
  - Noise, reverberation – multiple levels and types
- Intrinsic – voice is a performance
  - Demographic – sex, age, etc. – cross-speaker
  - Speech style – conversational, oratorical, read, etc.
    - Shouted and whispered speech
  - Aging
  - Health, illness – short or long term
  - Stress, vocal effort (Lombard effect)
    - Difficult to simulate real-word forensic situations
- Parametric
  - Duration – often limited in crime scene recordings
  - Language/Dialect – limits range of human phonetics expertise

Role of Evaluation

- Since 1996 NIST has organized regular evaluations of automatic systems
- Can estimate error rates for conditions represented in the data
- Used data collected and audited by the LDC
  - Largely conversational telephone speech
  - Also in-room interview speech with multiple microphone recording
  - Diversity in speakers, handsets, mics, duration
  - Some efforts at variation in language, dialect, noise, vocal effort
  - But what can be collected is limited by cost and collection practicalities

NIST HASR (Human Assisted Speaker Recognition) Tests

- Addition to 2010 and 2012 main evaluations of automatic systems
  - Systems could use human expertise perhaps in combination with automatic systems
  - Systems might utilize individual experts and/or panels of naïve listeners
  - Limited to 15-20 trials (HASR1) or 150-200 trials (HASR2)
    - Selected to be particularly difficult cross-channel subset of main evaluation trials
- Overall results were not impressive
  - Best results with humans did not outperform best automatic system results
  - Trial sizes were very limited and results should be interpreted with great caution

NIST Forensic Sciences

Organization of Scientific Area Committees (OSAC)

- Initiative of NIST and DoJ to establish collaborative partnership with forensic science community, in some cases superseding prior scientific working groups
- Digital/Multimedia SAC includes Speaker Recognition Subcommittee
- Initial meetings January 2015
- Speaker subcommittee is currently conducting virtual meetings with the aim of producing initial documents suggesting some guidelines and best practices all can agree upon – these are in initial stages

NIST Vendor SRE Evaluation

- Speaker Recognition Evaluation for Vendor Systems (SREVS)
- Will focus on speaker detection in context of conversational speech over telephony and microphone channels under realistic noise conditions
- Participants will provide executable systems that are potentially deployable in operational environments
- Sequestered data may reflect variability of vocal effort/ stress levels, transmission channels, microphones, noise levels, and collection conditions encountered in actual law enforcement scenarios
- Initial pilot may be conducted late in calendar year 2016 based on prior SRE data

Abstract

- Speaker recognition has a rather checkered history in terms of its use in a forensic context. Past claims of a capability to produce "voiceprints" that could be regarded as comparable to fingerprints were vastly inflated if not downright false, and provided source material for key U.S. court rulings on what constituted acceptable bases for scientifically acceptable forensic evidence. It must be recognized that we currently generally cannot assert with certainty, based on automatic methods or human expertise, that a given person is the speaker in a particular recording. There are many factors that may affect the quality of available audio evidence in terms of being able to reliably make match/non-match decisions between voices, and what gets recorded at typical crime scenes is likely to be particularly challenging in terms of the channel qualities of the recording media, the durations of the utterances, and the cooperativeness and the physical and emotional states of the people involved. NIST and other organizations have been involved in studying the capabilities of both automatic and human based systems in performing successful speaker recognition with respect to varying types of speech utterances and to underlying channel

Abstract (cont'd)

- and environmental conditions, but practical and ethical considerations make it exceedingly difficult to investigate the kinds of stressful conditions likely to prevail at crime scenes. Professional audio investigators need to recognize the limitations of what their methods can determine, and that in many circumstances it may be advisable to decline to pursue work on a specific case and "punt". Often they might be more effective in investigatory roles to consider possible leads and rule out possible suspects than as primary expert witnesses in court. In its most recent evaluations of speaker recognition technology, NIST has conducted small scale evaluations, on a limited number of difficult trials, of systems encompassing human experts, showing limited success compared to the best performing automatic systems. In 2015 NIST and the Department of Justice created a Speaker Recognition Committee as part of its Organization of Scientific Area Committees (OSAC), to collaborate on creating consensus documentary standards and guid