Issues, Controversies and Advancements in Forensic Speaker Recognition

Dr. Jim Wayman

San José State University

Organization of this Talk

- Historical roots and controversies
- Impact of forensic speaker recognition on US law
- Different methods of examination
- Automated methods: How they work
- U.S. government tests
- The current controversy over reporting
- Speaker recognition in the news
- Voice stress
- Privacy concerns
- Moving forward with standards and a Scientific Working Group

Historical Roots:The Sound Spectograph of WWII

W. Koenig, H. K. Dunn and L. Y. Lacy, Journal of the Acoustical Society of America18(1) , July, 1946

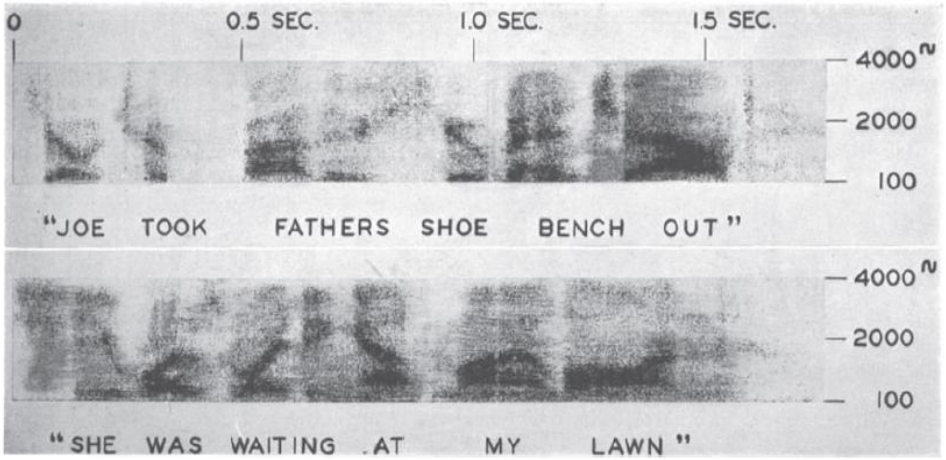


FIG. 8. Spectrograms made with the first laboratory model assembled from available equipment. The vocal resonances are clearly indicated. Further description of the features of spectrograms will be given in connection with subsequent illustrations.

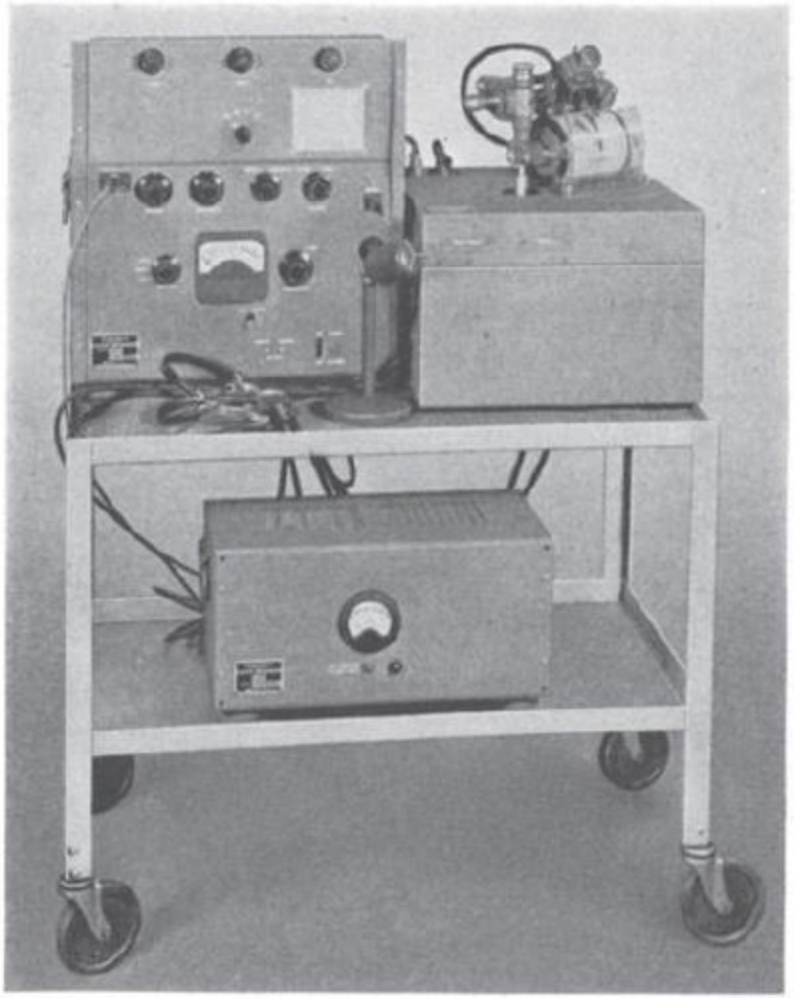


FIG. 9. The present model of the sound spectrograph, built in three parts for portability. The magnetic tape unit is in the right hand box, whose superstructure also carries the paper drum and stylus. The amplifiers, analyzer, etc., are in the left-hand box. The lower unit houses the regulated power supply.

Visible Speech (1942-2012)

San José State UNIVERSITY

- C. H. G. Gray and G. A. Kopp. “Voice Print Identification”, Internal Report. New York, NY: Bell Laboratories, 1944
- ‘Speech and Facsimile Scrambling and Decoding: A basic text on speech scrambling: Including the solution of speech privacy systems through the analysis of sound spectrograms”, Columbia University Division of War Research (1946). Under contract to the Office of Scientific Research and Development
- Potter, Kopp, Green, Visible Speech, Bell Labs (1947)
- J. Hershey, “Making Do”, Digital Signal Processing (1991)

WWII Scrambling and Decoding

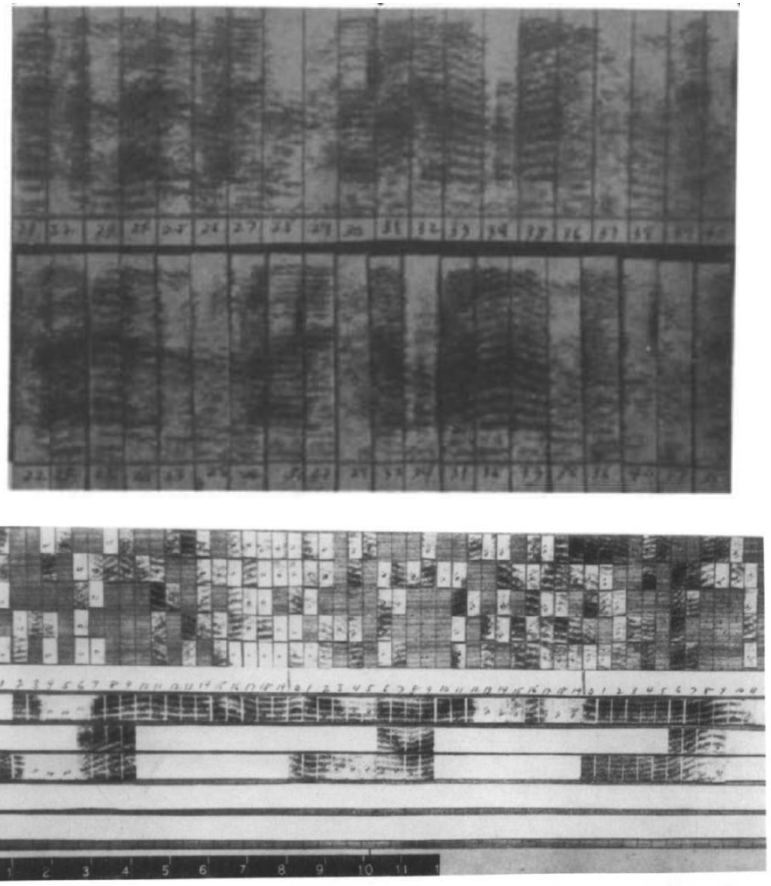


FIG. 6. Matching spectrograph patterns of two-dimensional scramble.

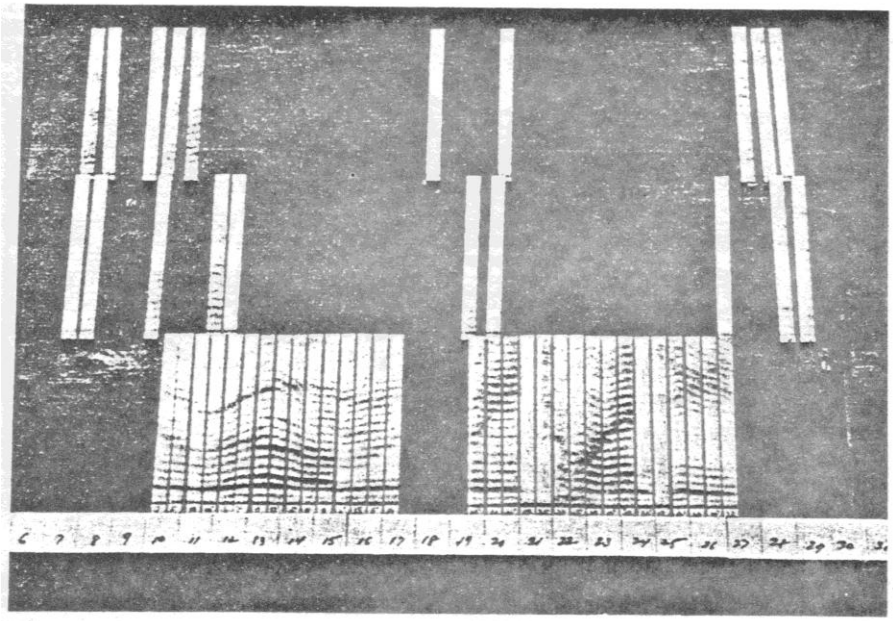


FIGURE 12. Method of matching spectrograph patterns of nonrepeated code TDS.

The "Voiceprint" Controversy

- "Closely analogous to fingerprint identification… uses the unique features found in their utterances….voice pattern uniqueness then rests on the improbability that two speakers would have vocal cavity dimensions and articulator use patterns identical enough to confound voiceprint …. methods" – L. Kersta, "Voiceprint Identification", Nature 4861, (1962)
- "…identification success in excess of 99%" -- L. Kersta, "Voiceprint Identification Infallibility", JASA 34(12) (1962)
- "…the available results are inadequate to establish the reliability of voice identification by spectrograms..many (scientists) are deeply concerned about the use of spectrographic evidence in the courts" -- R. Bolt, et al, "Identification of a Speaker by Speech Spectrograms", Science 166 (1969)
- R. Schwartz, "Voiceprints in the United States – Why they won't go away", P*roc. IAFPA (2006)*

Impact on US Law: US v. Smith, US Court of Appeals - 869 F.2d 348 (1989)

- "Tanya and Tamara Smith are identical twins who are commonly mistaken for one another….In their scheme, the two women posed as bank employees and telephoned banks authorizing them to make fictitious wire transfers of nonexistent funds… Because identity was a core dispute at trial, the government called a spectrographic voice identification expert to testify"
- "Smith also seizes on the fact that Nakasone admitted in his testimony that the field itself was controversial and that some studies had found high error rates."
- " Dr. Nakasone readily admitted that no one's voice is one-hundred percent unique, and that the field of voice identification is not one-hundred percent reliable."
- "We conclude that the district judge did not abuse his discretion in admitting Nakasone's testimony into evidence."

Daubert v. Merrill Dow Pharmaceutical, 509 U.S. 579 (1993)

Testimony is admissible as "scientific" if:

- Theory or technique has or can be tested
- Subjected to peer review and publication
- Existence and maintenance of standards for use (referencing U.S. v Williams)
- General acceptance in scientific community
- Known potential rate of error (referencing U.S. v Smith)

Different Methods of Forensic Examination

- Visible speech
- Phonetic
- Automated (computer) algorithms

Phonetic/Phonemic

Rose, Forensic Speaker Identification (2002)
F. Nolan, "Speaker identification evidence: its forms, limitations, and roles"(2001)

San José State UNIVERSITY

- Traditional phonetics, as in the International Phonetic Alphabet
- "Different pronunciations indicate different speakers *unless* explained by a coherent model of variation, principally a sociolinguistic one."
  - "Broad Australian" accent:
    - "Make" pronounced as "Mike"
    - "Hay" pronounced as "high"
- "machines and measurements are of little value without the complex process of interpretation which the phonetician brings to speaker identification"
- Clicks and non-speech vocal sounds

Advances in Fully Automated Systems

- S. Pruzansky, “Pattern-matching procedure for automatic talker recogntion”, JASA (26) pp. 403-406, 1963
- K.P. Li, et al, “Experimental studies in SV using an adaptive system”, JASA (40), pp.966-978, 1966
- K. Stevens, et al, “Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material”, JASA (44), pp. 1596-1607, 1968
- J.E. Luck, “Automatic Speaker Verification using Cepstral Measurements”, JASA, (46), pp. 1026-1031, 1969

Automated Approaches: HowThey Work

Extracting Cepstral Coefficients from telephone data

- Telephone bandwidth (300 => 2700 Hz)
- Digitize data (8k samples/sec)
- Frame and window data (256 samples using 128 new samples in each frame)
- Perform FFT on digitized data
  - 128 frequencies between 0 and 4000 Hz.
  - Bin width 31.25 Hz
- Find energy at each frequency

Cepstral Coefficients

- For Mel-scale, rescale frequencies in some way:
  - Example:
    - linear to 1000 Hz
    - logarithmic spacing above 1000Hz
  - 24 point filter bank
- Take log of interpolated energy in each frequency bin
- Take FFT (or cosine transform)

So What is a "Cepstrum"?

- "Spec" spelled backwards + trum
- The cepstrum represents the periodicity in the energy spectrum
  - Harmonically related sounds result in energy periodocity
- Why is this a good way to recognize speech or speakers?
- Bishnu Atal's challenge (1999)

Gaussian Mixture Models

- From F. Bimbot, et al, "A Tutorial on Text-Independent Speaker Verification", EURASIP Journal on Applied Signal Processing (2004)
- For a particular speaker, the probability afforded a cepstral feature vector x from a state i (i=1,…M) is

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{N/2}\left|\Sigma_i\right|^{1/2}} \; e^{-\frac{1}{2}\left((\vec{x}-\vec{\mu}_i)\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)^T\right)}$$

Why is Σ^{-1} Inside Distance Computation?

- Σ_i is covariance matrix of N-dimensional cepstral coefficients in a single state cluster i=1,…M
  - Real, symmetric (square) matrix
  - Orthogonal eigenvectors
  - V = matrix of eigenvectors
  - Λ = diagonal matrix of eigenvalues (same order as columns of V)
  - $WW = \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & \ddots & \\ 0 & & \ddots \end{bmatrix}$ $W = \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \sqrt{\lambda_2} & \\ & \ddots & \\ 0 & & \ddots \end{bmatrix}$

$$\Sigma_i = V_i \Lambda_i V_i^T = V_i (W_i W_i) V_i^T$$

To Invert Σ

$$\Sigma^{-1} = V\, W^{-1}\, W^{-1}\, V^{T}$$

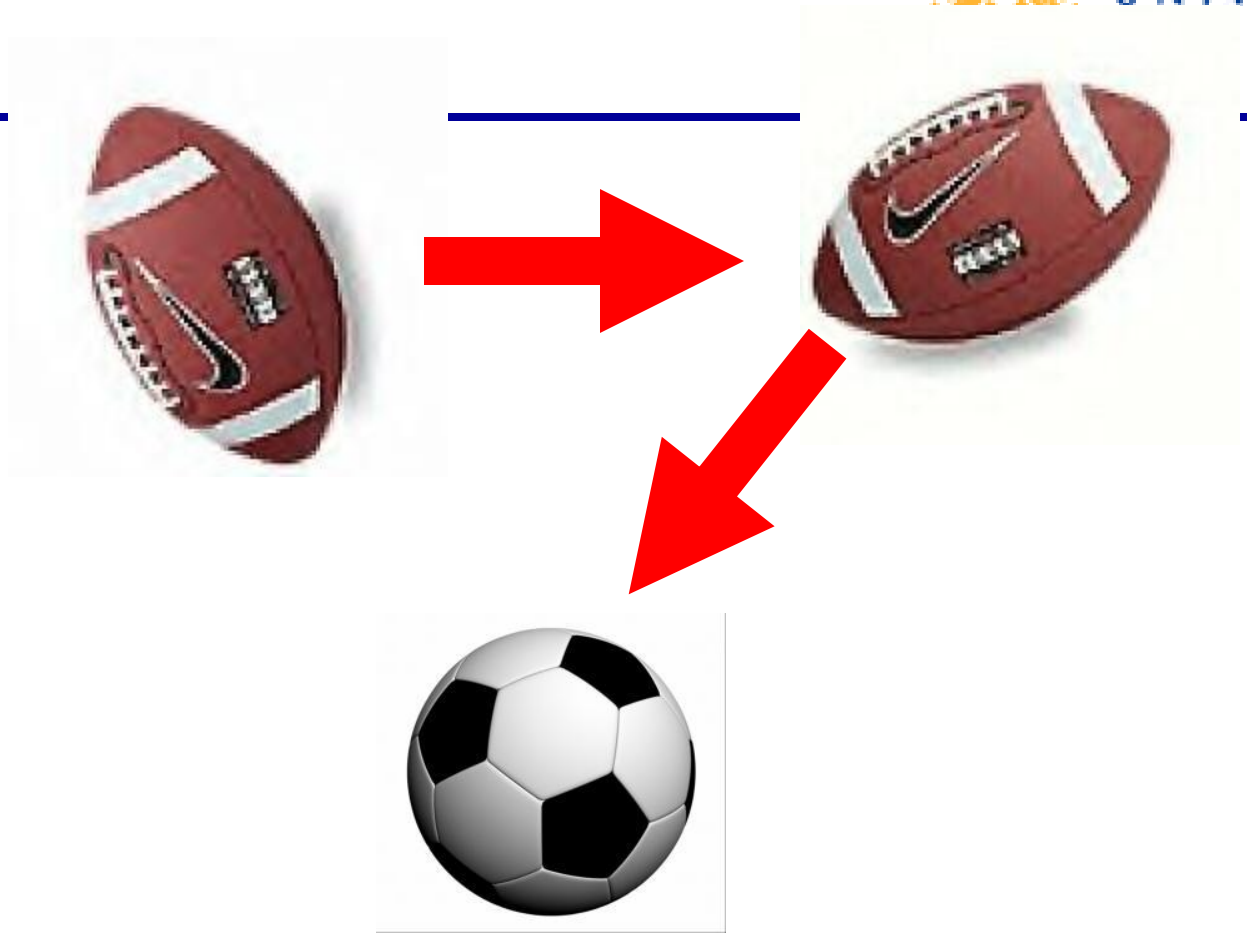
Proof:

$$\Sigma\, \Sigma^{-1} = V\, W\, W\, V^{T} V\, W^{-1}\, W^{-1}\, V^{T} = I$$

$$\left(\vec{X} - \vec{\mu}\right) V W^{-1} \left[\left(\vec{X} - \vec{\mu}\right) V W^{-1}\right]^{T} = \left(\vec{X} - \vec{\mu}\right) V W^{-1} W^{-1} V^{T} \left(\vec{X} - \vec{\mu}\right)^{T}$$

is the squared euclidean distance between the vector X and cluster mean, when the cluster is "whitened" by Σ^{-1} to make it round such that direction doesn't matter.

Rotate and Rescale



GMM

- But for each speaker, each state has a different *a priori* probability, p_i
- So the probability of getting x given a particular speaker S and M states is

$$p(x|S) = \sum^{M} p_i b_i(x)$$

GMM

If there are multiple cepstral vectors x_t over time

$$p(x|S) \propto \prod_{t=0}^{T} p(x_t|S) = \sum^{T} \log p(x_t|S)$$

Check $\frac{P(x|S)}{P(x|\sim S)}$ for each speaker against threshold

GMM

- At enrollment, for each speaker create distributions for M states (each with a mean and covariance matrix)
  - Seed with Universal Background Model (UBM)
- With unknown sample x, for each speaker compute

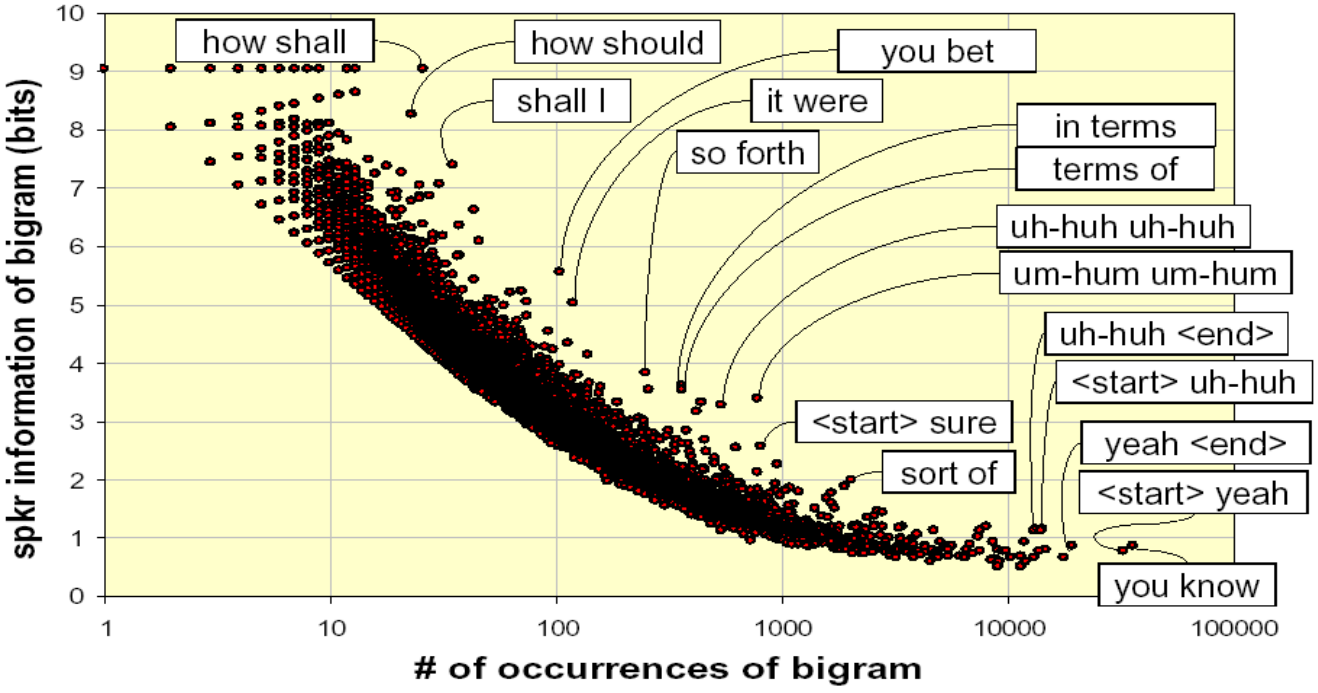
$$P(x|S) = \sum^{M} p_i b_i(x)$$

- With T samples over time, compute for each speaker

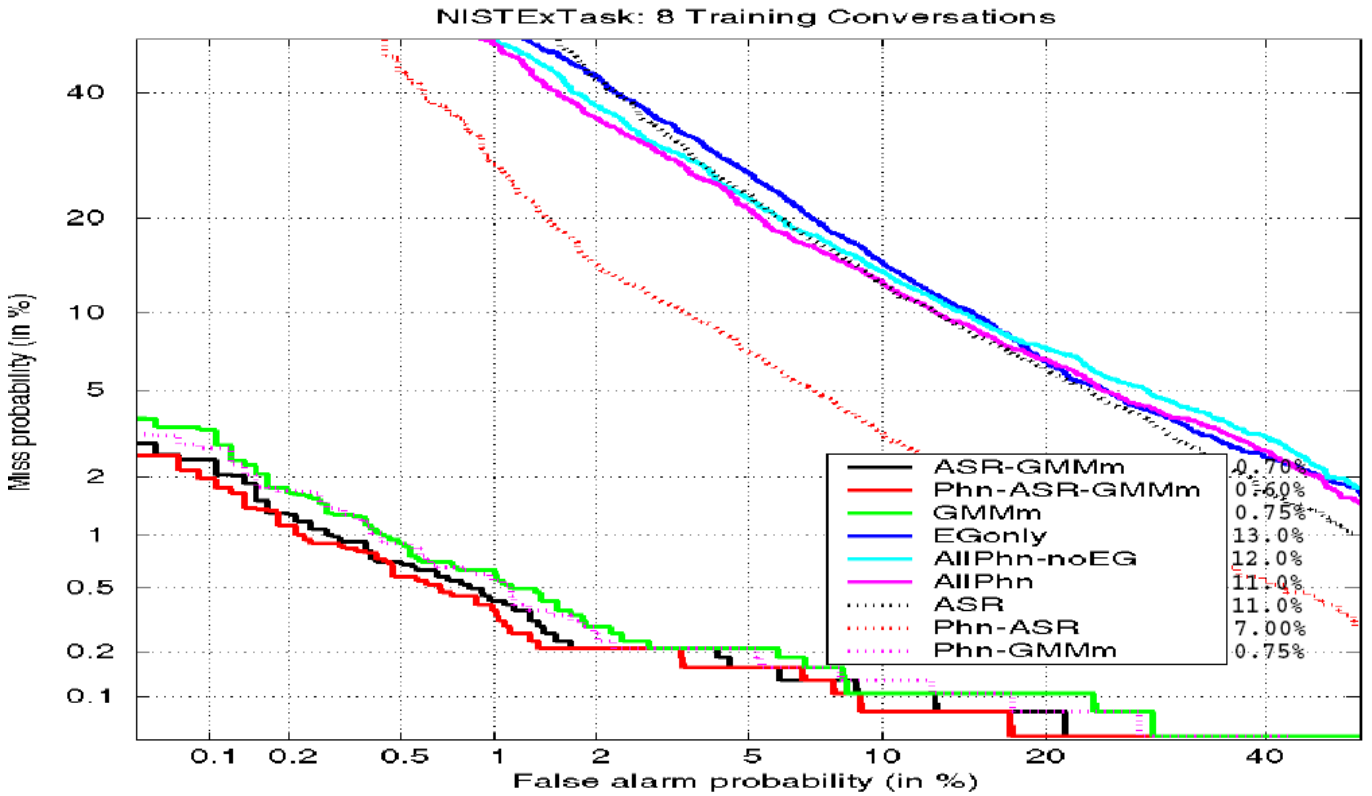
$$P(x|S) = \sum_{t=0}^{T} \log P_t(x|S)$$

- Compare each $\frac{P(x \mid S)}{P(x \mid \sim S)}$ to threshold (under assumption of equal Bayesian priors)

Higher-level Speaker Characteristics: Speech Bi-grams



Idiolectics

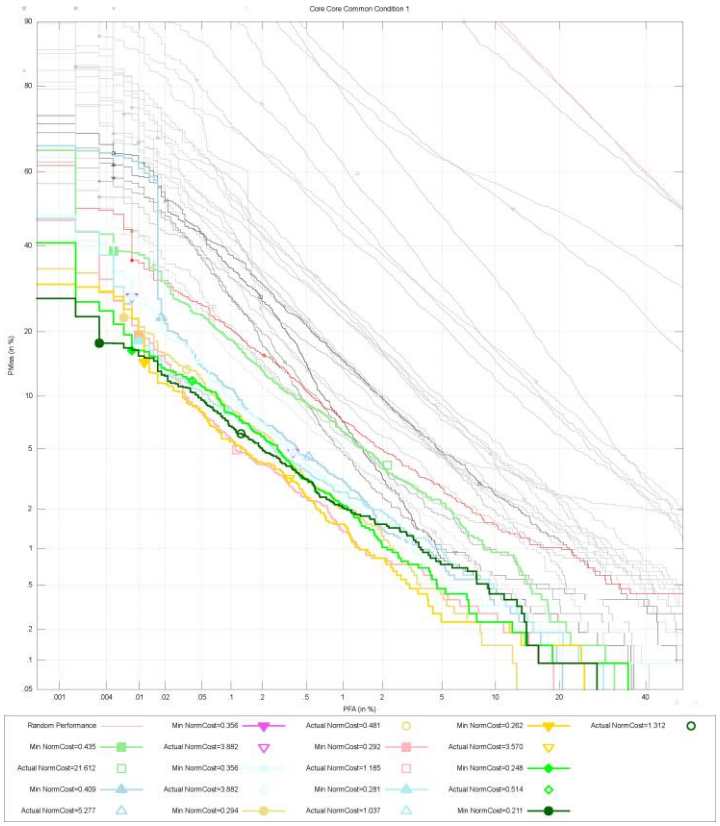


New Methods

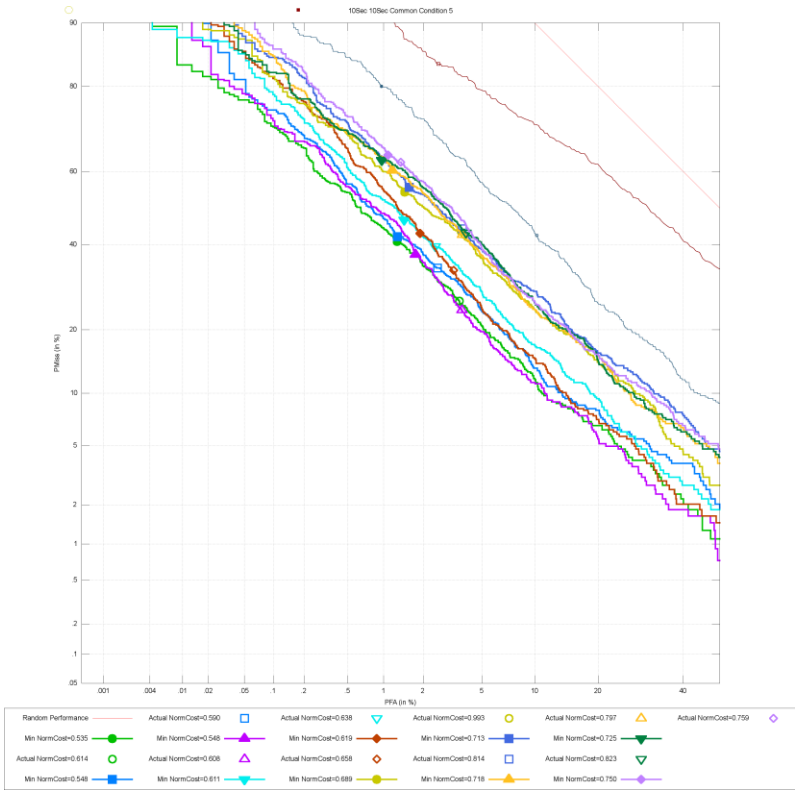
- Support Vector Machines
  - For each sample, find M N-dimensional cluster means
  - Concatenate means into super vector
  - Ignore Σ_i and state prior probabilities b_i
  - Computer euclidean distance in MxN space
- i-Vectors
  - Compute within-class and between-class Σ in MxN space
  - Linear discriminant analysis techniques

U.S. Government Tests: NIST Speaker Recognition Evaluation (SRE 1996-2012)

San José State University



SRE2010 5 min-5min, same mic



10sec-10sec telephone

NIST Human Assisted Speaker Recognition (HASR 2010)

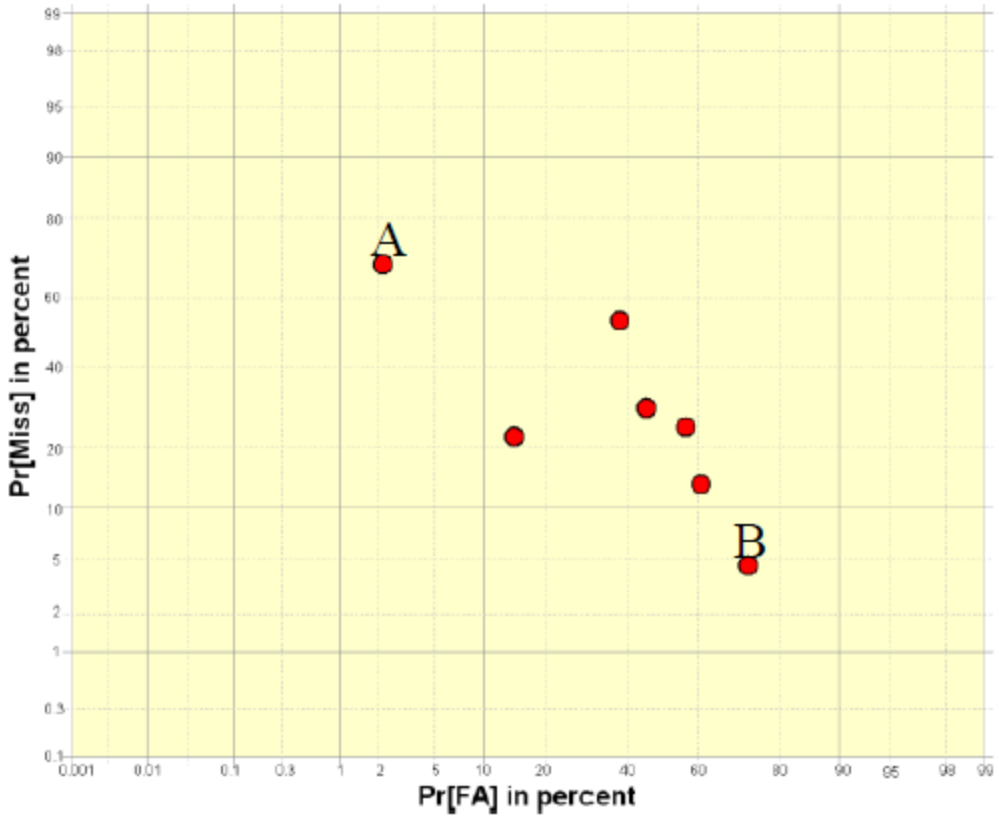


Figure 8: Numbers of misses and of false alarms of each of the 7 systems on HASR2 trials, plotted as DET points

SRE 2012 Participants

-Agnitio
-ANHYT
-atip Gmbh
-Auditech
-BAE Systems
-Beijing Institute of Technology
-Brno University of Technology
-CCNT Lab of Zhejiang University
-Clarkson University
-Cogent
-CpdQ
-CRIM
-Hong Kong Polytechnic University
-I4U (Institute of Infocomm Research / University of New South Wales/ University of Eastern Finland)
-IBM
-ICSI
-IDIAP
-Indian Institute of Technology, Guwahati
-Indian Institute of Technology, Hyderabad
-Indian Institute of Technology, Madras
-Institute of Information Science, Academia Sinica
-Institute of Acoustics, Chinese Academy of Science
-Institute of Infocomm Research
-Johns Hopkins University / MIT CSAIL / MIT Lincoln Labs
-LIMSI
-LRDE
-Multimedia Processing Lab (National Central University)
-Nanjing University
-National Taipei University of Technology
Nuance
-Ozyegin University
Politecnico di Torino
-Polytechnic University of Madrid
-Poznan University of Technology
-Queensland University of Technology / Universidad Autonoma de Madrid (ATVS)
-Radboud University, Nijmegen
-Sabanci University
-Shanghai Dragoon Voice
-Shenzhen Institute of Advanced Technology
-SRI
-ST Microelectronics
-STC-innovations Ltd
-Swansea University
-Tecnológico de Monterrey
-Tokyo Institute of Technology
-Tsinghua University (3)
-TUBITAK
-Universidad Politécnica de Madrid (UPM)
-University of Avignon
-University of Basque County
-University of Canberra
-University of Coimbra
-University of Eastern Finland
-University of Manchester
-University of Science and Technology of China
-University of Texas at Dallas
-University of Zaragoza
-Validsoft
-Vocapia
-Xiamen Talented Software

Current Controversies over Reporting: Likelihood Ratio

- "…(the) value (of speaker recognition evidence) is best expressed using a likelihood ratio. Referring to the definition of this likelihood ratio, the analysis of the scientific evidence does not allow the scientist alone to make an inference on the identity of the speaker" -- Christophe Champod, Didier Meuwly, "The Inference of Identity in Forensic Speaker Recognition", Speech Communication 31 (2000)

The Bayes Problem

- The inverse conditional probabilities

$$\frac{P(\textit{same source}\,|\,\textit{observed similarity})}{P(\textit{different source}\,|\,\textit{observed similarity})} = \frac{P(\textit{same source})}{P(\textit{different source})} \times \frac{P(\textit{observed similarity}\,|\,\textit{same source})}{P(\textit{observed similarity}\,|\,\textit{different source})}$$

- Option 1: Expert witness testifies to hard decision (i.e., same source or different source) with level of certainty
- Option 2: Expert testifies to left hand side above
- Option 3: Expert testifies to $\frac{P(\textit{observed similarity}\,|\,\textit{same source})}{P(\textit{observed similarity}\,|\,\textit{different source})}$

and jury adds their own priors $\frac{P(\textit{same source})}{P(\textit{different source})}$

Problems with the Likelihood Ratio

- "…the utility of mathematical methods for these purposes has been greatly exaggerated. Even if mathematical techniques could significantly enhance the accuracy of the trial process, …their inherent conflict with other important values would be too great to allow their general use."
- "Readily quantifiable factors are easier to process - and hence more likely to be recognized and then reflected in the outcome - than are factors that resist ready quantification. The result, despite what turns out to be a spurious appearance of accuracy and completeness, is likely to be significantly warped and hence highly suspect" -- Laurence H. Tribe, "Trial by Mathematics: Precision and Ritual in the Legal Process", Harvard Law Review, 84(6) (1971)
- "As phoneticians, we not in a position to use the Bayesian approach quantitatively (as)… say, in DNA evidence" – Nolan (2001)

Speaker Recognition in the News: The Trayvon Martin Case

- Feb. 26, 2012 shooting of 17-year old Trayvon Martin by 28-year old George Zimmerman
- Yelling and shooting recorded by police emergency call center
- With police, George Zimmerman reenacts scene
- Were the yelling voices on the two recordings from the same person?

FBI Report (2 April, 2012)

"Critical listening and digital signal analyses further revealed that the screaming voice of the 911 call is of insufficient voice quality and duration to conduct a meaningful voice comparison with any other voice samples primarily due to the screaming voice being: 1) produced under an extreme emotional state; 2) limited in the number of words and phrases uttered; 3) superimposed by other voices most of the time; and 4) distant, reverberant and very low signal level"

Voice Stress Analysis and Compensation

- Inherent challenges
  - Vague concept of "stress"
  - Laboratory experiments or reconstructions as proxies
    - Ethical constraints
  - Within speaker variability of response
  - Across speaker variability of response
- "Although the scientific community has concluded that (Voice Stress Analysis-) based technologies lack scientific validity, parts of the non-scientific community still believe in the value of these tools employing them in criminal investigation contexts." – Kirchhübel, et al (2011)

Further Reading on Voice Stress

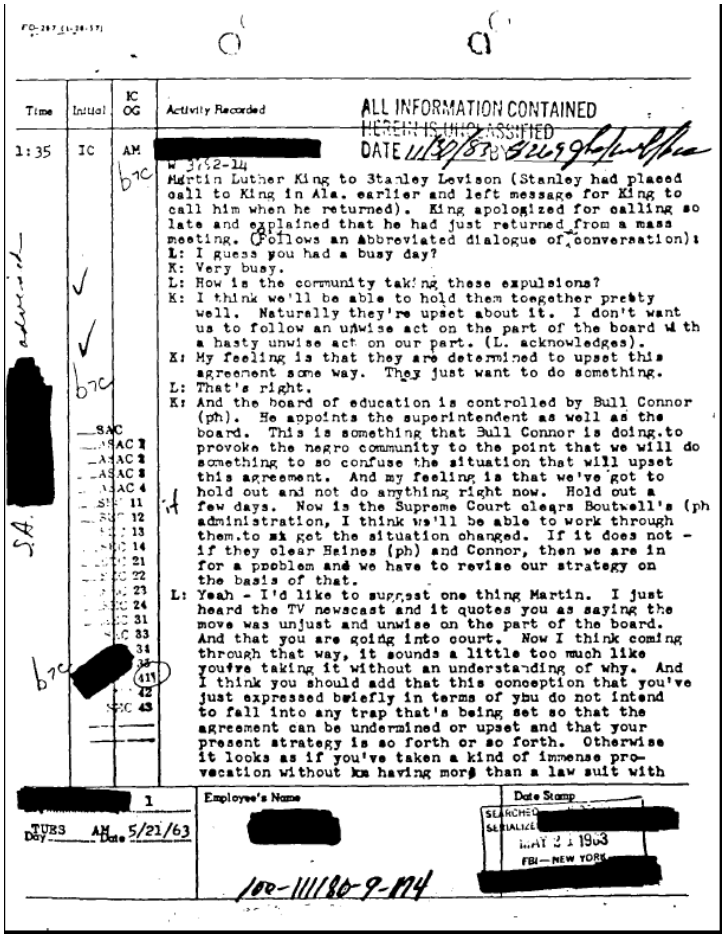
- C. Kirchhübel, et al, "Acoustic correlates of speech when under stress: Research, methods and future directions", Int. Journ. of Speech, Lang. & Law, 18(1), (2011)
- A. Eriksson and F. Lacerda, "Charlatanry in forensic speech science: A problem to be taken seriously", Int. Journ. of Speech, Lang.& Law, 14(2), (2007) (WITHDRAWN)
- K. R. Damphouse, "Voice Stress Analysis: Only 15 Percent of Lies About Drug Use Detected in Field Test", NIJ Journal 259 ()
- D. Haddad, et al, "Investigation and Evaluation of Voice Stress Analysis Technology: Final Report", Report to NIJ from Air Force Research Lab - Information Directorate, Doc. 193832 (2002)

Privacy Concerns

"This was a new science – finding a criminal by a print of his voice. Until now they had been identified by fingerprints. They called it dactyloscopy, study of the finger whorls. It had been worked out over the centuries. The new science could be called voice studies….or phonoscopy" -- Aleksandr Solzhenitsyn, "The First Circle" (1968)

FBI Wiretap Transcript of Martin Luther King, Jr. (5/21/'63)

www.lexisnexis.com/academic/1univ/hist/aa/content-d1.asp



Time | Initial | IC OG | Activity Recorded

ALL INFORMATION CONTAINED HEREIN IS UNCLASSIFIED DATE 11/29/83

1:35 | IC | AM

[redacted]

W 3762-14

Martin Luther King to Stanley Levison (Stanley had placed call to King in Ala. earlier and left message for King to call him when he returned). King apologized for calling so late and explained that he had just returned from a mass meeting. (Follows an abbreviated dialogue of conversation):

L: I guess you had a busy day?
K: Very busy.
L: How is the community taking these expulsions?
K: I think we'll be able to hold them together pretty well. Naturally they're upset about it. I don't want us to follow an unwise act on the part of the board with a hasty unwise act on our part. (L. acknowledges).
K: My feeling is that they are determined to upset this agreement some way. They just want to do something.
L: That's right.
K: And the board of education is controlled by Bull Connor (ph). He appoints the superintendent as well as the board. This is something that Bull Connor is doing to provoke the negro community to the point that we will do something to so confuse the situation that will upset this agreement. And my feeling is that we've got to hold out and not do anything right now. Hold out a few days. Now is the Supreme Court clears Boutwell's (ph) administration, I think we'll be able to work through them to get the situation changed. If it does not - if they clear Haines (ph) and Connor, then we are in for a problem and we have to revise our strategy on the basis of that.
L: Yeah - I'd like to suggest one thing Martin. I just heard the TV newscast and it quotes you as saying the move was unjust and unwise on the part of the board. And that you are going into court. Now I think coming through that way, it sounds a little too much like you're taking it without an understanding of why. And I think you should add that this conception that you've just expressed briefly in terms of you do not intend to fall into any trap that's being set so that the agreement can be undermined or upset and that your present strategy is so forth or so forth. Otherwise it looks as if you've taken a kind of immense provocation without [redacted] having more than a law suit with

[redacted] 1 | Employee's Name [redacted] | Date Stamp

TUES Day AM Date 5/21/63

SEARCHED SERIALIZED MAY 21 1963 FBI—NEW YORK

100-111180-9-114

Pub. Law 03-414 (1994) Communications Assistance for Law Enforcement Act (CALEA)

- ..isolating and enabling the Government, pursuant to a court order or other lawful authorization, to intercept all of the subscriber's wire and electronic communications…
- Requires the governmental entity to offer specific and articulable facts showing that there are reasonable grounds to believe that the contents of a wire or electronic communication… sought by the entity are relevant and material to an ongoing criminal investigation before a court order may be issued for disclosure of such information.
- Prohibits the use, production, or possession of an altered telecommunication instrument… to obtain unauthorized access to telecommunications services. Imposes 15 years' imprisonment and a fine of $50,000 or twice the value obtained by the offense.

Moving Forward: Voice Format Standards

- ISO/IEC JTC1 SC37 19794-13 Working Draft 3
  - Single speaker voice data for access control applications
  - Threat of cancellation due to lack of progress
- ANSI/NIST ITL 1-2 Type 11 Voice Record
  - In development
  - Wide variety of proposed use cases
  - Multiple speakers assumed
  - 3 public meetings to date

Scientific Working Group for Forensic and Investigatory Speaker Recognition

San José State UNIVERSITY

- US Department of Justice Sponsorship
- First meeting to be held March, 2013 time frame
- To establish “best practices” for collection, examination and reporting
- Other SWGs

SWGFAST	**SWGMAT**
SWGDAM	**SWGGSR**
SWGGUN	**SWGSTAIN**
SWGFEX	**SWGDVI**
FISWG	**SWGMDI**
SWG DOC	**SWGDOG**
SWGANTH	**SWGDRUG**
SWGTOX	**SWGDE**
SWGWILD	**SWGIT**

Concluding Remarks

- Speaker recognition is oldest automated human recognition technology
- Great practical need for progress as a forensic method
  - Standards for examination and reporting
  - Laboratory and examiner accreditation standards
  - Standards for collection & transmission of data and metadata
- Progress is being made:
  - ANSI/NIST Type-11 voice record
  - SWG-SPEAKER
  - NIST SRE and HASR programs