On Forensic Speaker Recognition Case Pre-Assessment

Gheorghe Pop
National Institute of Forensic Expertise
Bucharest, Romania

Dragoş Drăghicescu
Dragoş Burileanu
Speech and Dialogue (SpeeD) Laboratory
Faculty of Electronics, Telecommunications and IT
University "Politehnica" of Bucharest, Romania

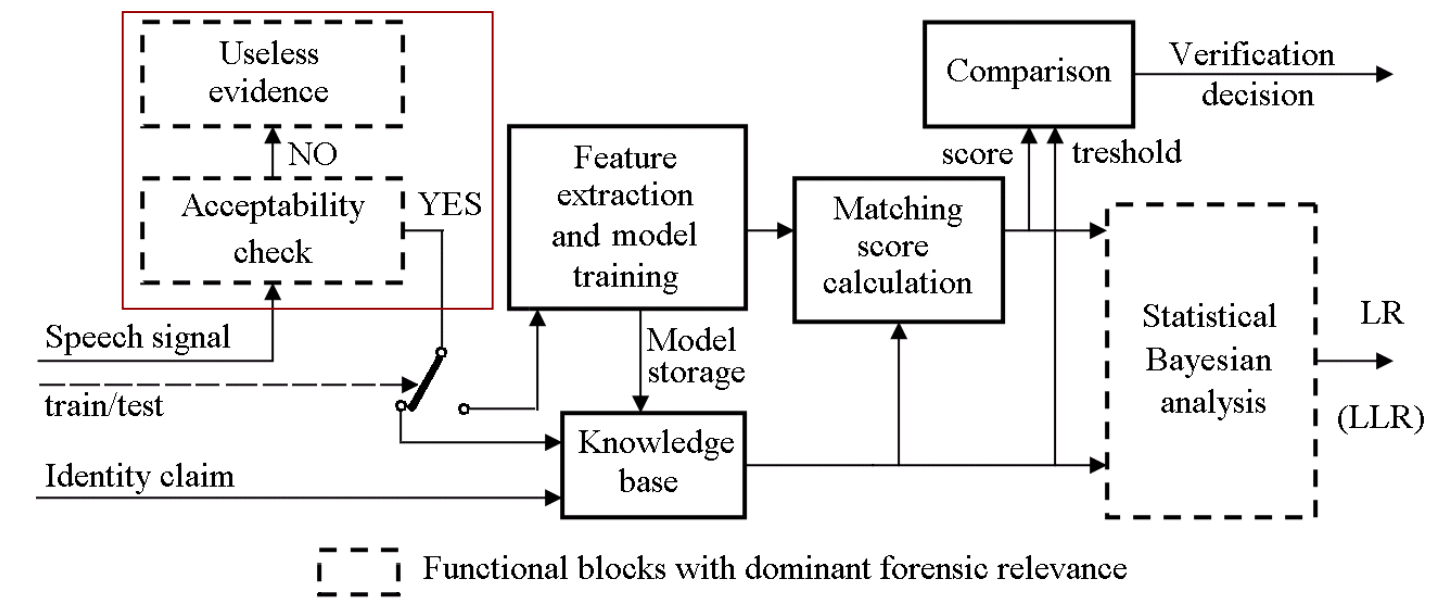
Outline of Speaker Recognition Case Assessment

- Early forensic audio techniques
  - difficult to explain in court
  - acceptance and rejection received in their original domains
  - both analysis techniques and evidence filtered by court debate
- Actually needed
  - coherent approaches including case assessment
  - balanced, logical, transparent and robust
  - easy to explain
- Interpretation of evidence before analysis
  - case pre-assessment
  - preliminary classification of evidence based on relevance
  - objective criteria based on quality parameters of the speech

Early Techniques

- Old paradigm (speakers are uniquely identifiable by their speech)
- Labs were equipped with oscilloscopes, analog filters, variable speed players, microscopes, sound spectrographs
- No answer on:
  - What are the potential findings
  - What is their potential value
  - What are the probabilities of these findings under the competing hypotheses

Forensic Speaker Recognition Systems



- In forensic speaker recognition, target speaker is first detected, then evidence is assessed based on his speech quality
- Both biometric and forensic speaker recognition use speech quality assessment, but their use is not the same:
  - Biometric systems are open to ask for repeats
  - Forensic systems work offline, with speech recorded elsewhere

Techniques for Defining and Assessing Speech Quality

- Subjective vs. objective
- Online vs. offline
- Telecom standards
- Conditions from dedicated software
- Limits on individual parameters of speech
- Limits from the perceptual domain
- Techniques based on combined effects of factors relevant to recognition

Channel Factors Adverse to Speaker Recognition

- Mutes
- Sharp declines
- Signal interruptions
- Unnatural silence or beeps
- Robotization
- Frame repeats
- Noise of all sort

Usual Minimum Quality Limits Imposed by Recognition Software

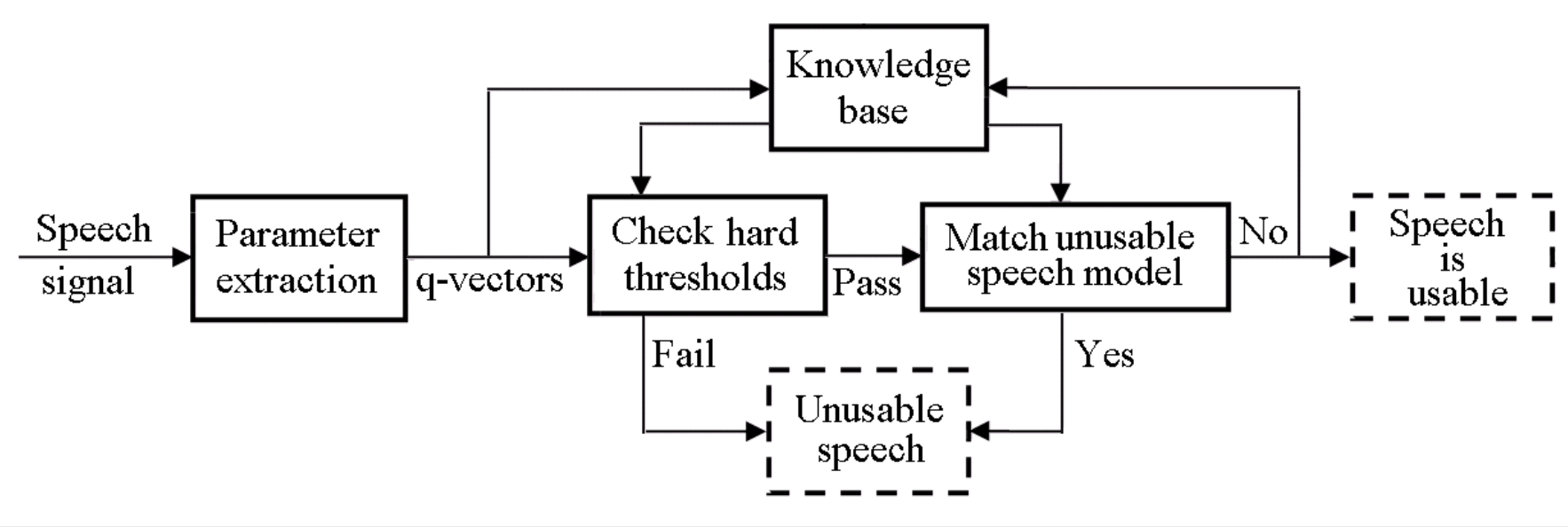
- Matched conditions vs. mismatched
- Enough pure speech for recognition (>16 s)
- Clipping less than 25 %
- Global SNR > 10 dB
- Reverberation time < 300 ms
- Flat frequency response of the channel at least from 300 Hz to 3400 Hz

Objective Speech Quality Evaluation from Telecommunication

- Intrusive methods (double ended)
  - PESQ (ITU-T P.862)
  - PEAQ (ITU-R BS.1387)
- Non-intrusive methods (single ended)
  - PESQ (ITU-T P.563)
  - LCQA [27]

A speech quality standard in forensic speaker recognition has not been established yet

Forensic Speaker Recognition
Case Pre-Assessment



Speech quality parameters were extracted by short time analysis, and used as q-vectors to form a model of unusable speech

If quality parameters fall under any hard threshold, speech is unusable

For speech with quality superior to this lowest level, quality is assessed based on results from previous recognitions that failed because of low quality signal

Proposed Speech Quality Evaluation

- Automatic, based on short time analysis
  - Apply hard local thresholds on speech parameters
    - Clipping, drops, artifacts, T60 / ALC, MOS score
    - Speaker detection, SNR, non-linearity (HOS)
  - Check if passed speech matches a model of unacceptable speech
    - Irrelevant speech from previous examinations was used to model unacceptable speech (q-vectors)
- Experimental setup
  - On a selected speech database, recognition scores by VoiceNet system were collected
  - Correlation was searched between these parameters and recognition scores

Experimental Database

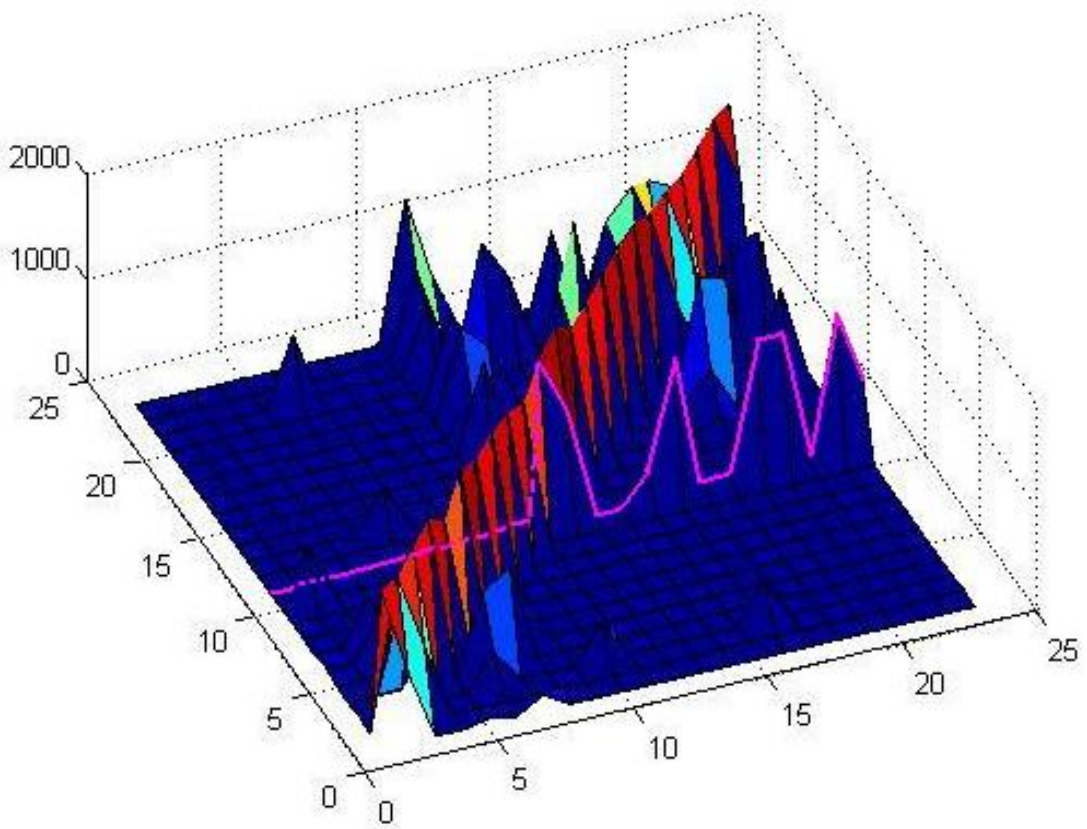
ELSDSR, from IMM, Denmark [35]:

- Read speech database created in 2006
- Designed for speaker recognition systems development and evaluation
- Speech from 10 female and 13 male speakers
- Audio recordings last from 3.5 to 17 seconds
- Same recording equipment and location
- There are seven train recordings and two test recordings for every speaker in the database

ELSDSR – English Language Speech Database for Speaker Recognition

IMM – Institute for Mathematical Modeling

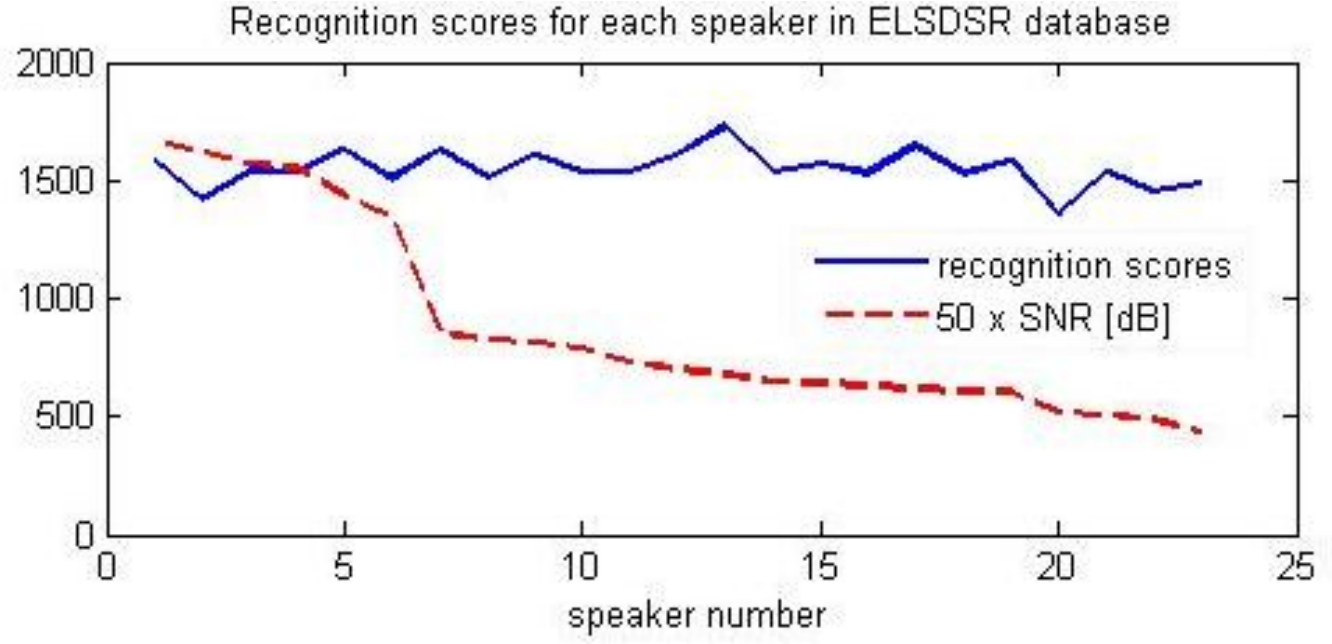
Confusion Matrix from STC VoiceNet Software on ELSDSR



Recognition scores of speaker MASM in the confusion matrix

Hard Thresholds at Minimum Levels of Speech Quality Parameters

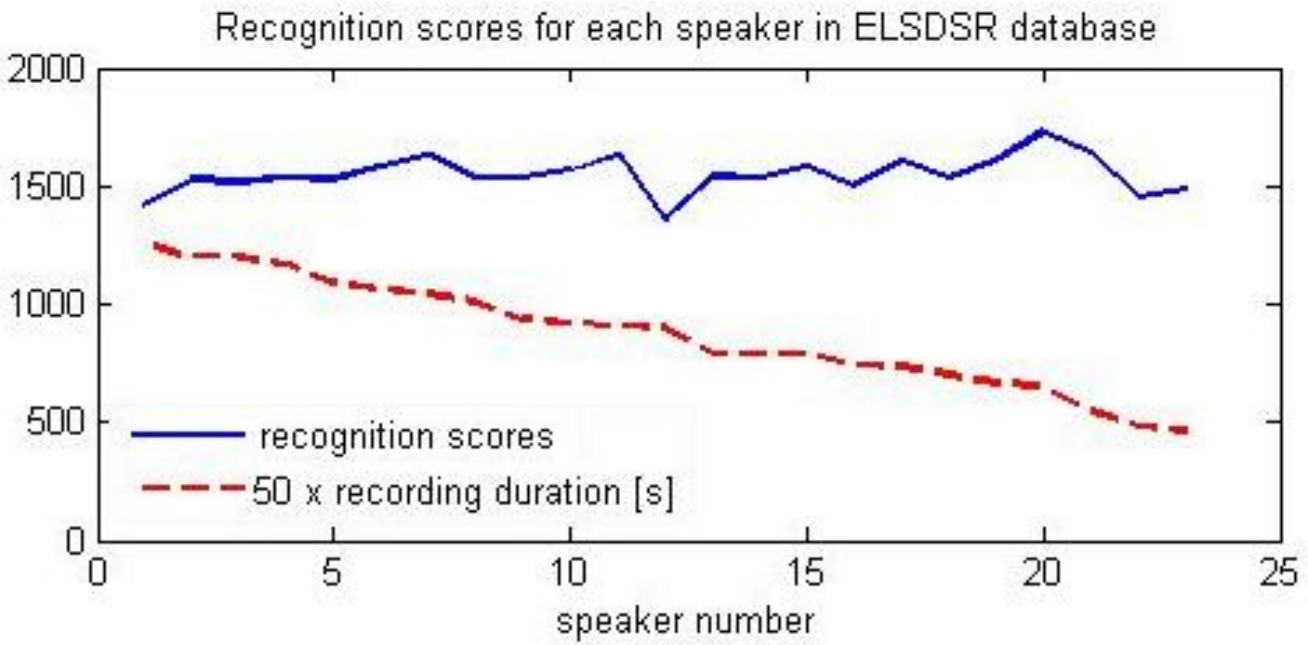
Required level of global SNR (at least 10 dB)



SNR values of ELSDSR speech recordings are from 10 dB to 32 dB

Hard Thresholds at Minimum Levels of Speech Quality Parameters

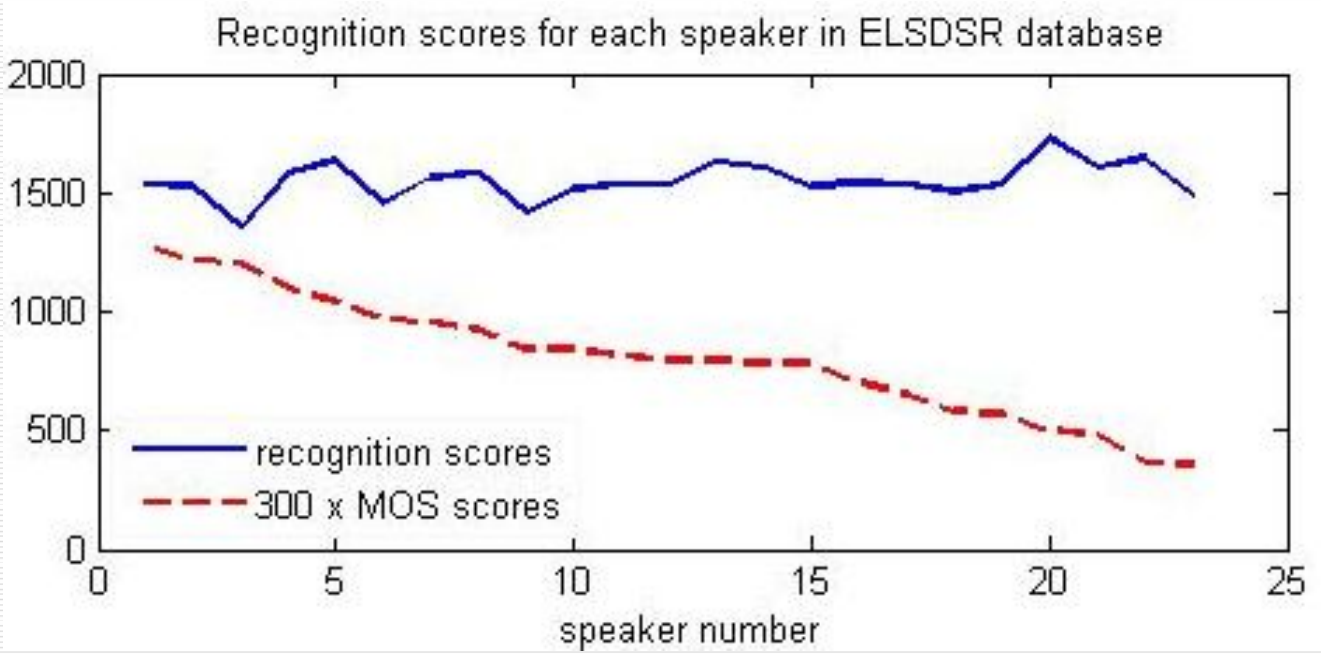
Required duration of pure voice (at least 16 s)



Pure voice durations values of ELSDSR speech recordings are from 5 s to 17 s

Hard Thresholds at Minimum Levels of Speech Quality Parameters

Required level of MOS score (at least 2.0)



MOS scores for ELSDSR speech recordings are from 1.3 to 4.3

Advantages of the Procedure

- Long high quality recordings may be locally unusable just over the target speech
- Long low quality recordings may be high quality just over the target speech
- Regions in the speech signal with high distortion effects can be removed so as to avoid their use in speaker recognition
- Same goes for badly located mutes and interruptions as well as for other periods with strong channel effects

What Next?

- Speaker evidence with parameter levels less than minimum get rejected
- There still is a chance that some passed recordings are irrelevant
- Improved pre-assessment by using q-vectors (signal quality parameter sets as vectors)
- Evidential recordings can be automatically analyzed, with better selection of material relevant to forensic speaker recognition

On Forensic Speaker Recognition Case Pre-Assessment