UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

An overview of Speaker Verification technologies applied to telephony services

Dr. Jean Hennebert
Maître Assistant – Multimedia Engineering DIVA
Computer Science Department
Université de Fribourg

Plan

An overview of Speaker Verification (SV) technologies applied to telephony services

- A bit of introduction
- Particularities of the speech signal
- What identifies a speaker?
- Pro and cons of SV
- Algorithms Fundamentals of SV
- Performances of state-of-the-art SV systems
- Taxonomy of SV systems
- Applications in telephony services

- Conclusions - What's next?

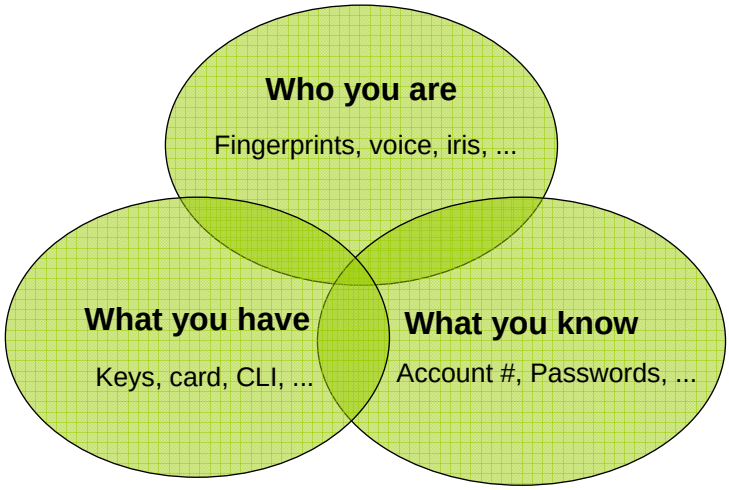
UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

A bit of introduction

SV is a biometric technology
Kinds of biometrics
The goal of the game

jean.hennebert@unifr.ch, University of Fribourg

SV is a biometry



jean.hennebert@unifr.ch, University of Fribourg

Kinds of biometrics

- Fingerprint,iris
- Speech
- Signature

Physical attributes

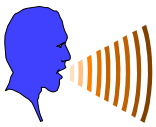
Performance attributes

Rigid / passive

Plastic / dynamic

jean.hennebert@unifr.ch, University of Fribourg

The goal of the game



+ claimed identity

Speaker Verification = Authenticate someone's claimed identity on the basis of her / his voice

SCORE

Threshold Value

"True speaker": access granted

"Impostor": access denied

jean.hennebert@unifr.ch, University of Fribourg

UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

Particularities of the speech signal

Difficulties of the speech signal

The main difficulty is...

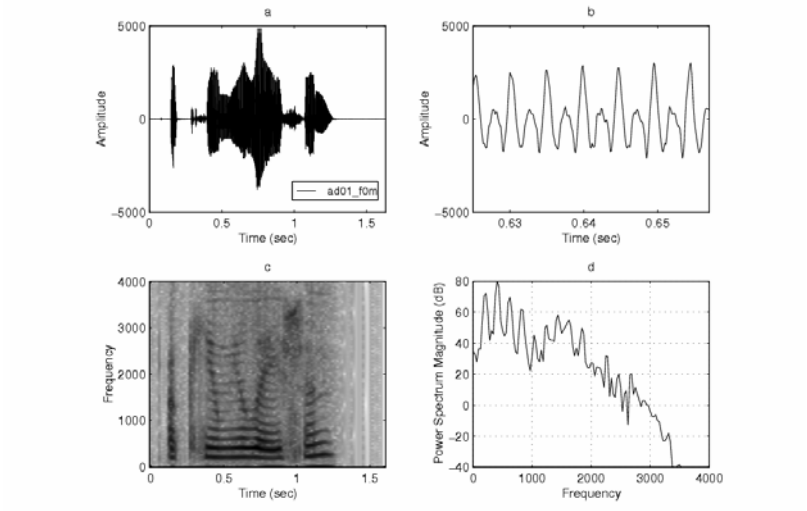


Figure 1.1: Speech signal of the word *accumulation* (a) : waveform, (b) partial waveform, (c) narrowband spectrogram of (a), (d) power spectrum magnitude of (b).

Speech Signal Difficulties in Telephony Environment

- Due to telephony channel
  - limited bandwidth, channel
  - variability
  - environmental noise
- Due to service constraints
  - Speaker independence
  - Barge-In capability
- Due to the language
  - homophone, ambiguities
  - coarticulation
- Due to the user
  - Next slide!

The main difficulty is...

- Not used to speak to a computer
- High expectations
- Easily frustrated
- No discipline
- Short-term auditive memory
- Hesitations, fillers, breathings
- Phone from any place
- Poor language skills
- Technology rejection

→ **Users should be educated and motivated**

UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

What identifies a speaker ?

3 sources of variation among speakers
Current SV algorithms capture a part of it

jean.hennebert@unifr.ch, University of Fribourg

3 sources of variation among speakers

1. Physiological properties
   - Vocal tract shape
   - Vocal cords length
2. Behavioral characteristics
   - Speaking rate
   - Prosody
   - Coarticulation
3. Higher level information
   - Vocabulary selection
   - Grammatical constructions
   - All sort of hesitation and filler sounds

jean.hennebert@unifr.ch, University of Fribourg

Current SV algorithms capture a part of it

1. **Physiological properties**
   - Vocal tract shape
   - Vocal cords length
2. **Behavioral characteristics**
   - Speaking rate
   - Prosody
   - Coarticulation
3. Higher level information
   - Vocabulary selection
   - Grammatical constructions
   - All sort of hesitation and filler sounds
   - Conversation context

UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

Pro and cons of SV

3 Pros
3 Cons

3 pros

1. Good user acceptance
   - Considered as lowly intrusive
   - Talking is a natural gesture
   - No physical contact with the sensor
2. Low technology cost
   - Simple microphone can be used
   - Potentially over any telephone
3. Pretty good security against impostor attacks
   - Challenge-response strategy can be used
   - Imitations capture the behavioral characteristics, not the physiological ones

3 cons

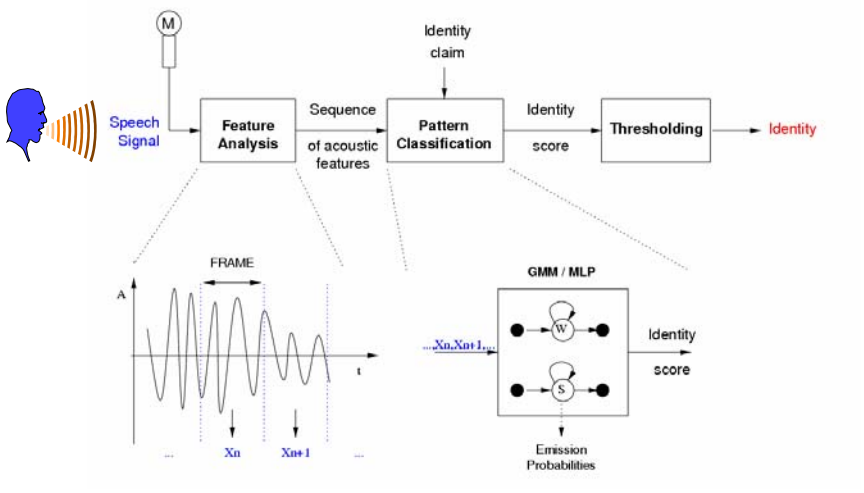
1. Enrollment session
   - One session does not capture all of the variabilities
   - Incremental enrollment may be necessary
   - The shortest is the best for the user
   - The longest is the best for the technology
   - You need to secure the enrolment
2. Medium accuracy
   - Usually less ranked as other biometrics such as fingerprints or iris scan
   - Variability is the cause!
   - Uniqueness is not so good (family members)
3. Pretty bad security against impostor attacks
   - If not properly designed, a simple recording of the user's voice can break into the system

UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

Algorithms fundamentals

Overview
Detection problem
Threshold setting
ROC / DET curve

jean.hennebert@unifr.ch, University of Fribourg



$$S_c = p(X|M_{client}) = \prod_{n=1}^{N} p(x_n|M_{client})$$

$$S_w = p(X|M_{world}) = \prod_{n=1}^{N} p(x_n|M_{world})$$

jean.hennebert@unifr.ch, University of Fribourg

SV = detection problem

- SV only gives you a **SCORE** ("log-likelihood ratio")

$$R_c = \log(S_c) - \log(S_w)$$

  - The decision is taken according to a threshold value *T*

$$R_c > T \quad \rightarrow accept$$
$$R_c \leq T \quad \rightarrow reject$$

- **Detection problem**: 2 types of errors:

1. False Acceptation (FA) ("*false alarm*")
2. False Rejection (FR) ("*missed detection*")

The rate of FA and FR depends to the threshold value

Threshold setting

For a given threshold value *T:*

- The system shows a probability of falsely rejecting a client

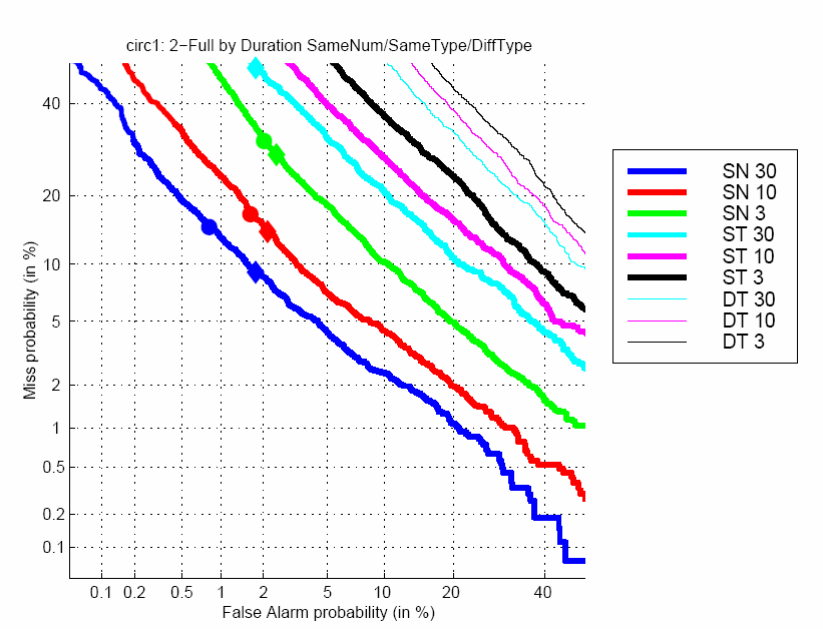
$$P(reject|client)$$

- The system shows a probability of falsely accepting an impostor

$$P(accept|\overline{client})$$

➡ You have to determine a level of security for your system ! This is usually done minimizing a cost function:

$$C_{det} = C_{fr}P(reject|client)P(client) + C_{fa}P(accept|\overline{client})P(\overline{client})$$



jean.hennebert@unifr.ch, University of Fribourg

UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

Performances of state-of-the-art SV systems

Performance is a function of ...
Some figures
Ways to break into the system

jean.hennebert@unifr.ch, University of Fribourg

Performance is a function of ...

- **Amount of data**
  - enrollment / test duration
  - multiple enrollment session
- **Quality of speech signal**
  - channel: bandwidth, recording device, reverberation...
  - background noise, multiple sources of speech
- **Modeling strategy**
- **User**
  - Level of cooperation
  - Health/stress state
  - Intrinsically, performances are not the same for different users
- **Time of the day** !!!

Some figures

- 50%
  - Upper bound limit of SV performance
- EER < 0.5%
  - What the technology vendors claim
- EER ~ 1%
  - What can be reasonably expected in a properly designed real-life telephone application
- EER ~ 20%
  - Few data for enrollment, few data for testing, noisy environment, mismatched conditions over the telephone

NIST organizes a “competition” of SV every year, open to research institutes and industries

Ways to break into a SV system

1. Pre-recording the client's voice
   - The recording has to be good quality because systems are generally sensitive to channel and recording conditions
   - Properly designed challenge-response systems prevent from such attacks
2. Forcing the client to give its sample
   - May not be working properly since stress impacts on the voice characteristics
3. Imitating the client's voice
   - May not work : imitations capture the behavioral characteristics, not the physiological ones
4. Building a text-to-speech synthesizer of the client's voice
   - This may work but it costs a lot.

UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

Taxonomy of SV systems

Text dependent
Text independent
Text prompted

Text Dependent

- **System selected password**
  - A priori fixed phrases, PIN
  - Identity claim and SV can be done at the same time
- **User selected password**
  - Technology is much more difficult
  - Recovery infrastructure

Text Independent

- User is free to say anything he wants
- User is not constrained to remember anything
- More vulnerability (any recording can be used to break into the system)

1. Text-Prompted

- Challenge-response
- User just has to repeat something prompted (easier for user and computer)
- System must check what has been said in a first step
- Randomness in the prompts prevents use of recorded speech

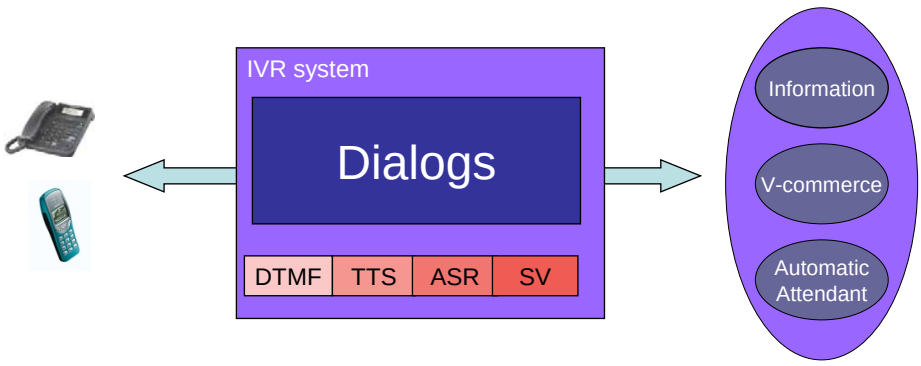
UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

Application in telephony services

Dialog machines
Some applications
Advantages for phonebanking
In practice...

jean.hennebert@unifr.ch, University of Fribourg

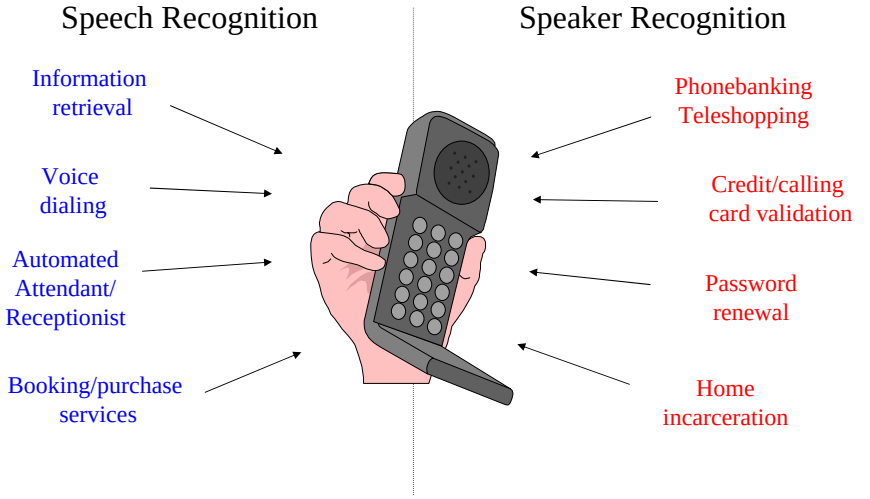
Dialog Machine



IVR = Interactive Voice Response
DTMF = Dual Tone Multiple Frequency
TTS = Text-to-Speech
ASR = Automatic Speech Recognition
SV = Speaker Verification

jean.hennebert@unifr.ch, University of Fribourg

Telephony commercial applications



In a banking environment: what are the advantages?

- SV is a biometrics: verify **who you are.**
- SV is user convenient: reduce need for PIN / Strike lists
- SV does not require elaborate installation/hardware on the user side.
- SV can be used as a Gate Keeper or Alarm Bell.
- SV can be a complement to *additional* security measures to be applied (passwords, strike lists, other biometrics,…).

UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

Conclusions

Critical Success Factors

- *Cooperation* with customer absolutely *necessary*.
- Technology must make life *easier* and *safer*.
- *Manage risk:* assess risk of every possible user activity.
- *Combine* other protection techniques with SV.

Conclusions

- Speech technologies will be used in many applications.
- Dialog systems with voice recognition applications are around the corner.
- SV will be available very soon for deployment:
  - there is a **good potential** for SV.
  - arguments are both **security** and **ease of use.**
  - technology is **continuously improved.**

What's next from a research point of view?

- **Modeling higher-level sources of information**
  - Longer term features
  - Weighting differently phonemes contributions
  - "going beyond the atomic units"
- **Multimodality**
  - Speech and face biometrics : talking faces
  - Speech and handwriting: S-SHARP