Speaker Recognition: Preliminary Requirements for CATS

IETF 53
21 March 2002

Larry Heck & Dan Burnett
Nuance Communications

Why am I up here?

User authentication

is usually the first step

for any user-focused human-computer spoken dialog.

Speaker Recognition
Outline

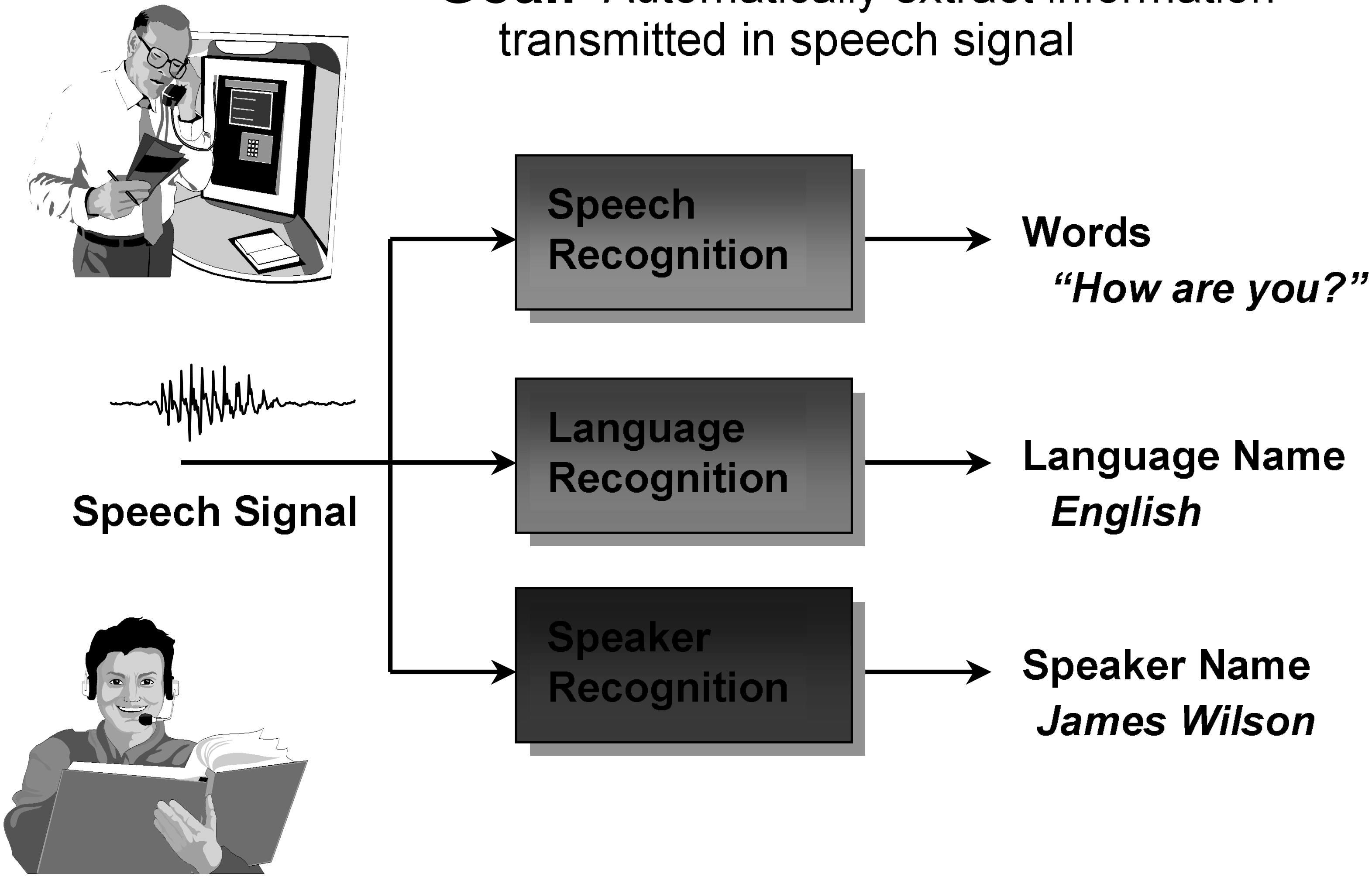
- **Overview of area**
  - Introduction/Terminology
  - General Theory
- **Needed application functionality**
  - Motivation: dialog design
  - Multi-utterance verification
  - Simultaneous identity claim and verification
  - Simultaneous knowledge and identity verification
- **Requirements**
  - Support for simultaneous ASR/verification/identification
  - Support integrated ASR/verification/identification
  - Support identified (named) resources

Speaker recognition
Outline

- **Overview of area**
  - Introduction/Terminology
  - General Theory
- **Needed application functionality**
  - Motivation: dialog design
  - Multi-utterance verification
  - Simultaneous identity claim and verification
  - Simultaneous knowledge and identity verification
- **Requirements**
  - Support for simultaneous ASR/verification/identification
  - Support integrated ASR/verification/identification
  - Support identified (named) resources

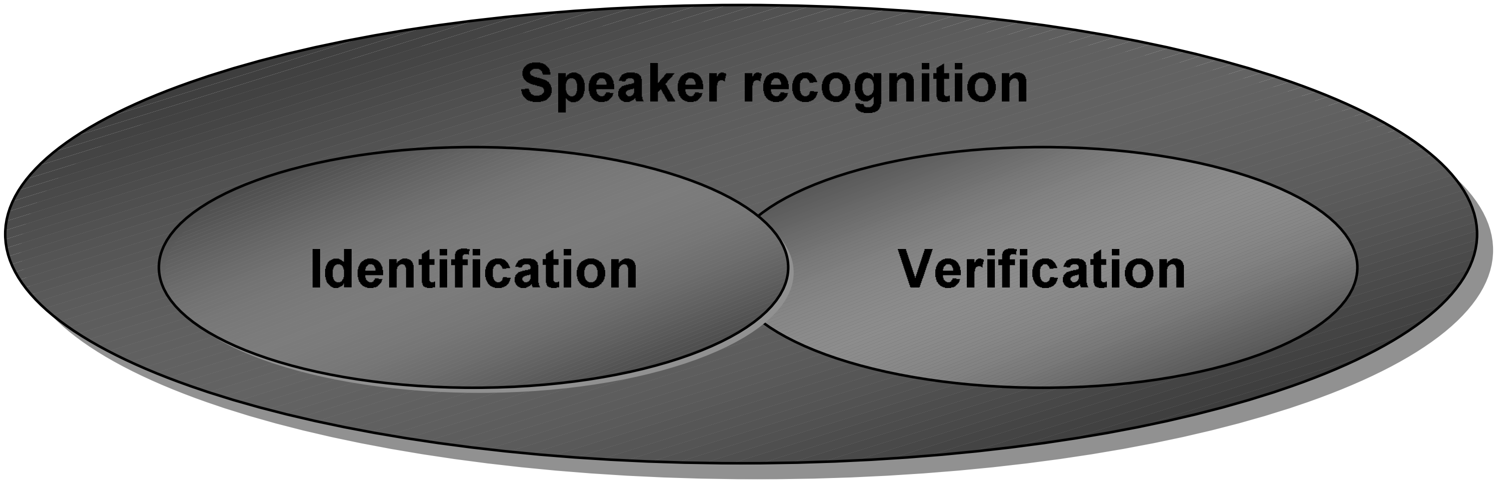
Extracting Information from Speech

Goal: Automatically extract information transmitted in speech signal



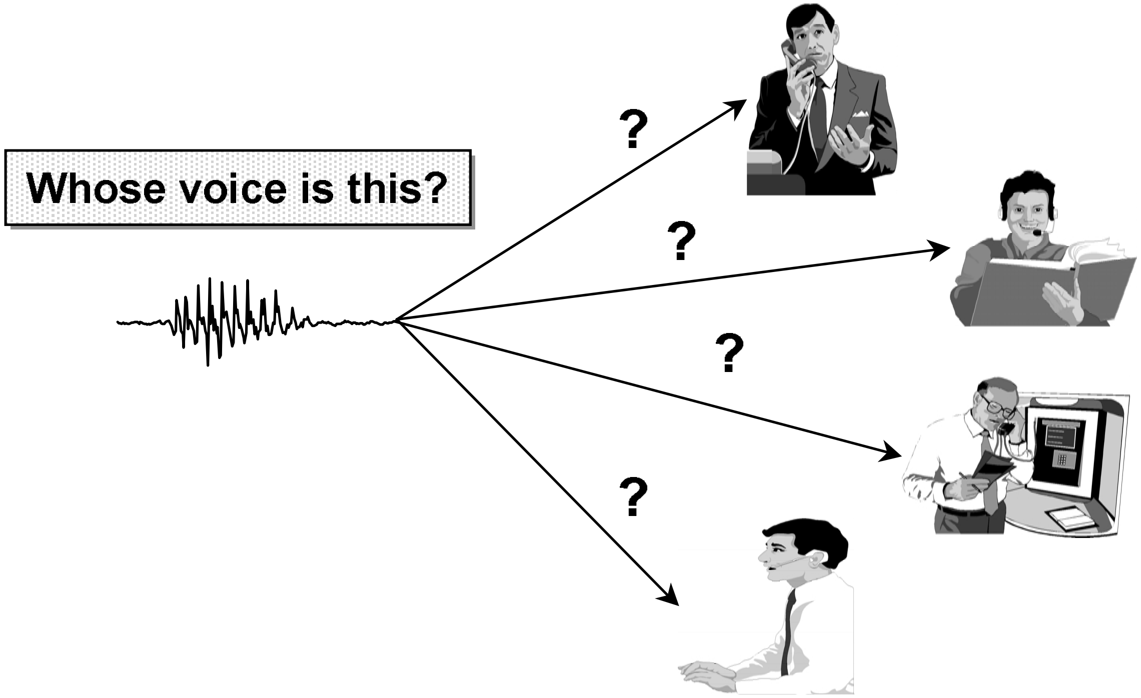
Terminology

- The general area of speaker recognition can be divided into two fundamental tasks



Terminology
Identification

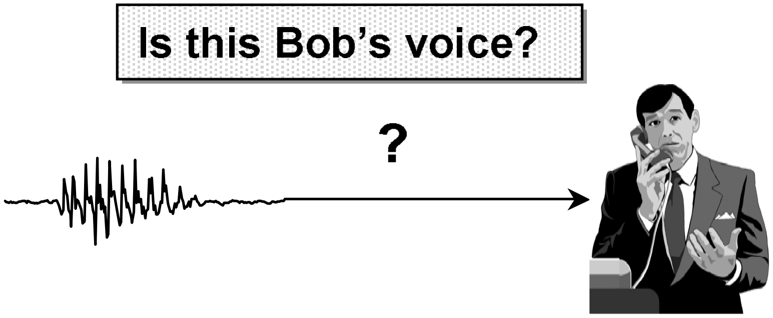
- Determines whom is talking from set of known voices
- No identity claim from user (one to many mapping)
- Often assumed that unknown voice must come from set of known speakers - referred to as closed-set identification



Whose voice is this?

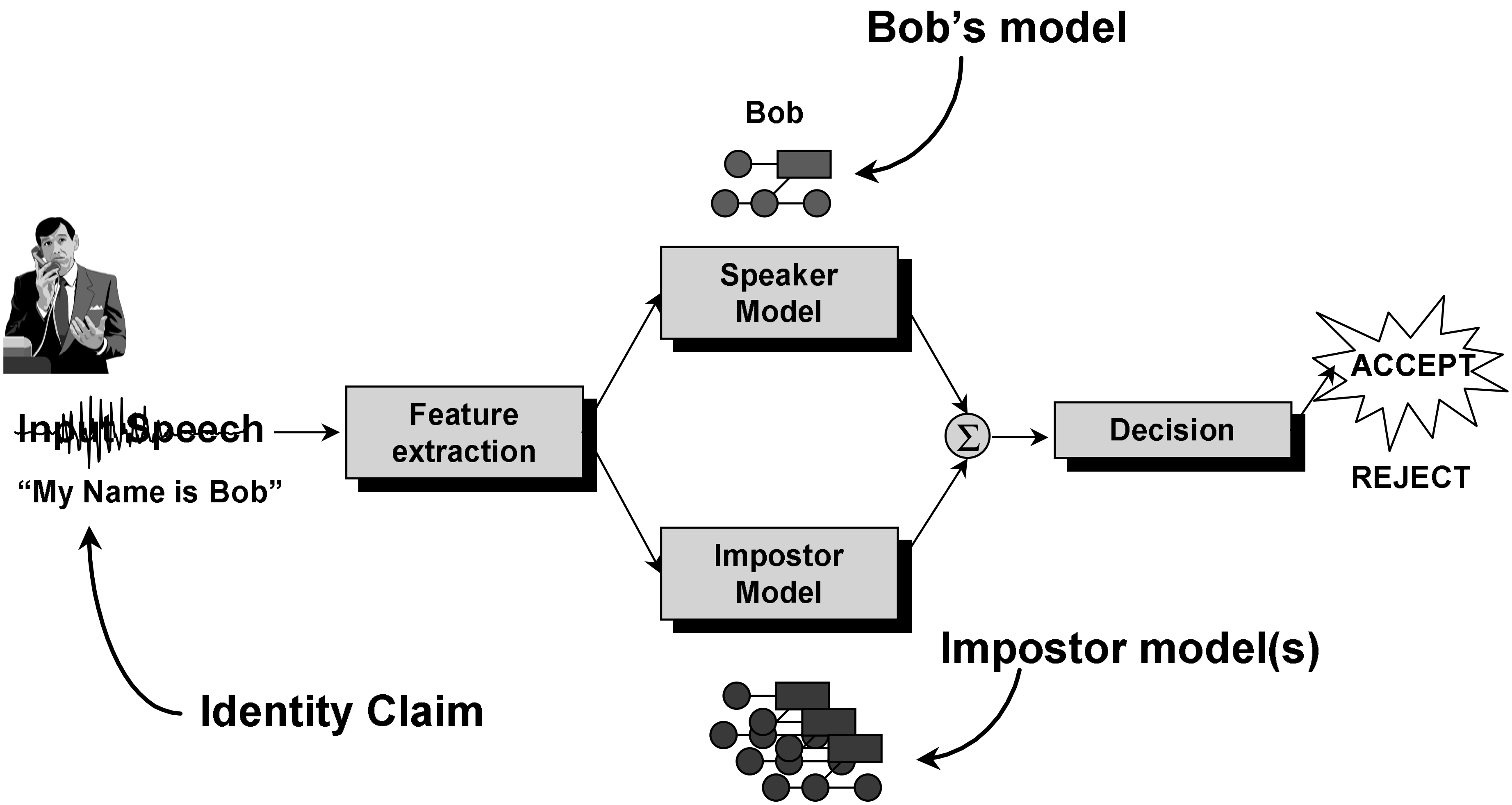
Terminology
Verification/Authentication/Detection

- Determine whether person is who they claim to be
- User makes identity claim (one to one mapping)
- Unknown voice could come from large set of unknown speakers - referred to as open-set verification



Is this Bob’s voice?

General Theory
Components of Speaker Verification System



Bob's model

Bob

Speaker Model

Input Speech

"My Name is Bob"

Feature extraction

Σ

Decision

ACCEPT

REJECT

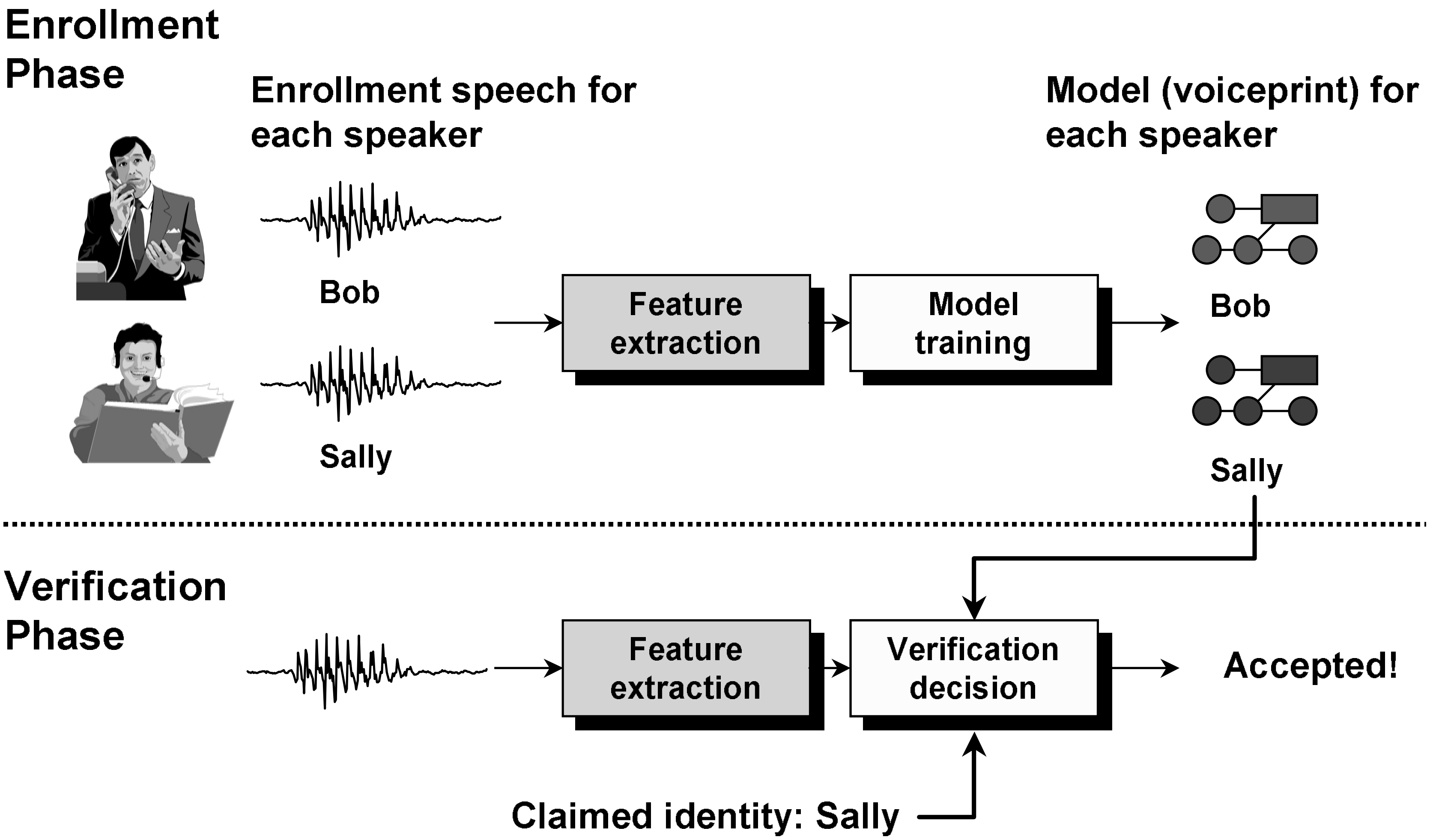
Impostor Model

Impostor model(s)

Identity Claim

General Theory
Phases of Speaker Verification System

Two distinct phases to any speaker verification system



Enrollment Phase

Enrollment speech for each speaker

Bob

Sally

Feature extraction

Model training

Model (voiceprint) for each speaker

Bob

Sally

Verification Phase

Feature extraction

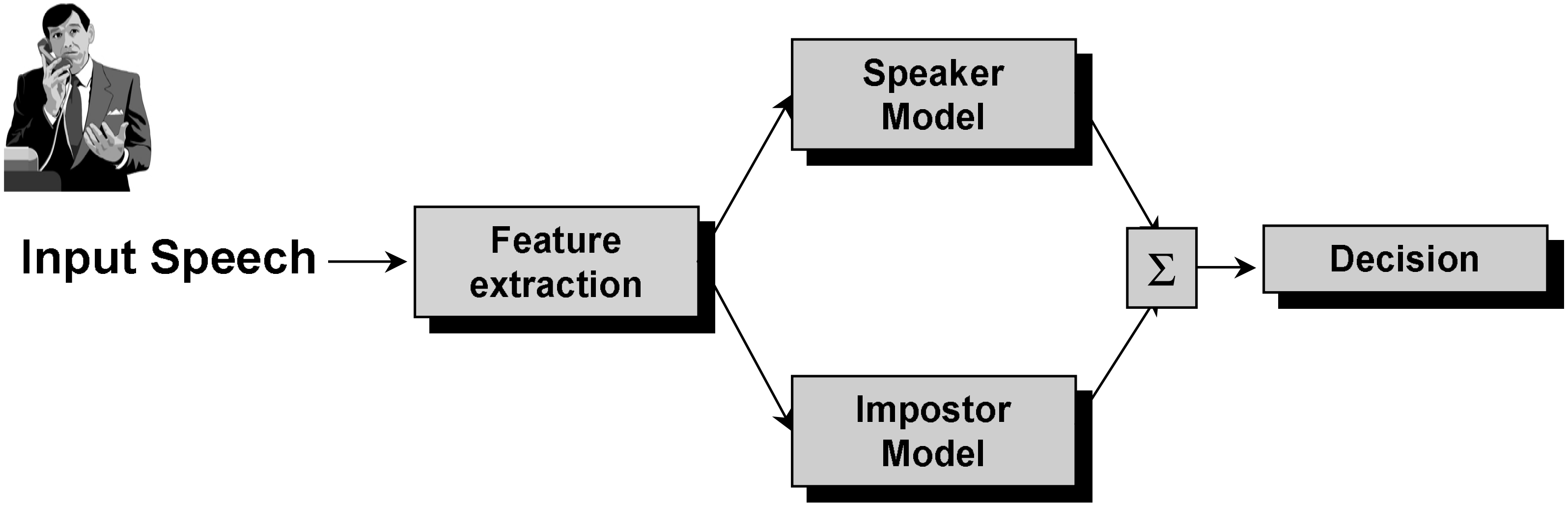
Verification decision

Accepted!

Claimed identity: Sally

General Theory

Components of Speaker Verification System



Input Speech

Feature extraction

Speaker Model

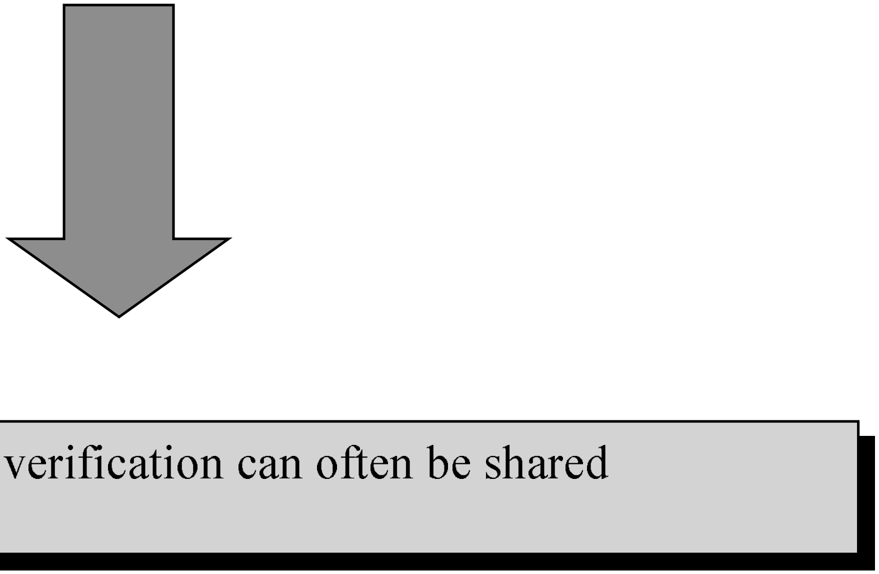
Impostor Model

Σ

Decision

General Theory
Features for Speaker Recognition

- Features derived from spectrum of speech have proven to be the most effective in automatic systems
- ASR uses similar computation and similar features



Computation for ASR and verification can often be shared

Speaker Recognition
Outline

- **Overview of area**
  - Introduction/Terminology
  - General Theory
- **Needed application functionality**
  - Motivation: dialog design
  - Multi-utterance verification
  - Simultaneous identity claim and verification
  - Simultaneous knowledge and identity verification
- **Requirements**
  - Support for simultaneous ASR/verification/identification
  - Support integrated ASR/verification/identification
  - Support identified (named) resources

Functionality Motivation
Spoken Dialog Design Principles

- **Dialog should be designed to be secure *and* convenient**
  - Security often compromised by users if dialog not convenient

    Example: 4-digit PIN

    Security = 1 out of 10,000 false accepts? NO!

    Users compromise security of PINs to make them easier to remember (writing down in wallet, on-line, etc.)

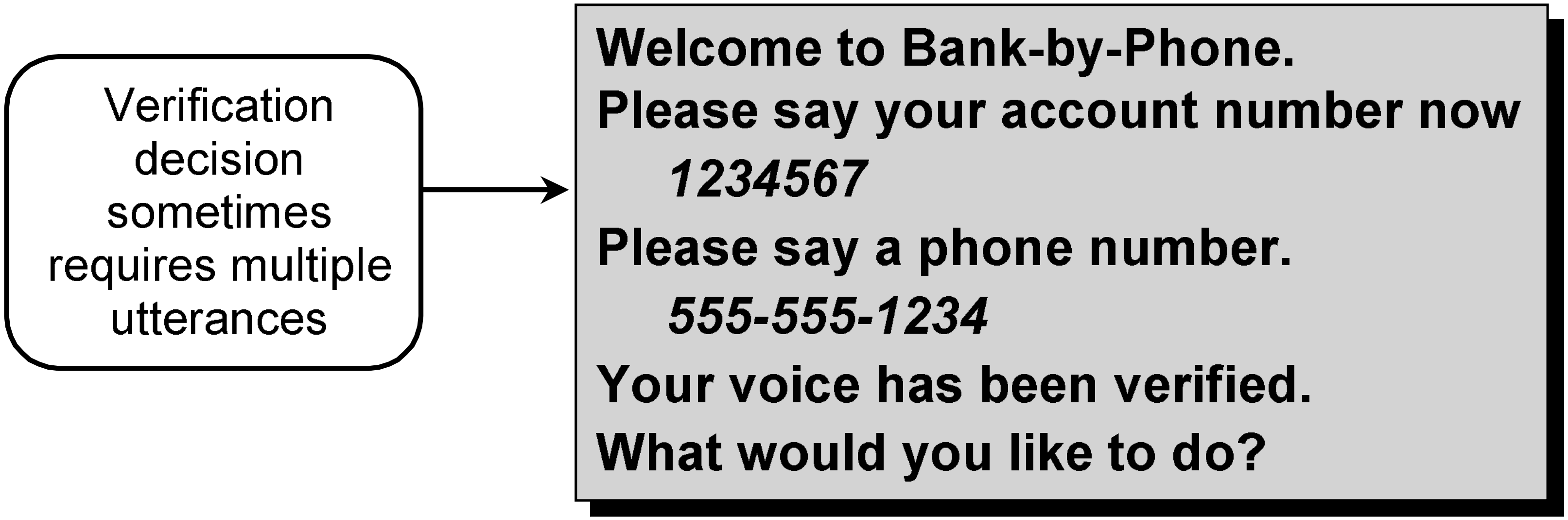
- **Dialog should be maximally constrained but flexible**
  - More constraints ➔ better accuracy for fixed length training
  - Example: balance between constraints on acoustic space while maintaining flexibility ➔ digit sequences

<u>Dialog Design Goal</u> ***Constrained but flexible dialog to maximize security while maintaining convenience***

Functionality

Default: Multi-utterance verification (1)

- **Verification accuracy improves with more user utterances.**



Verification decision sometimes requires multiple utterances

Welcome to Bank-by-Phone.

Please say your account number now

1234567

Please say a phone number.

555-555-1234

Your voice has been verified.

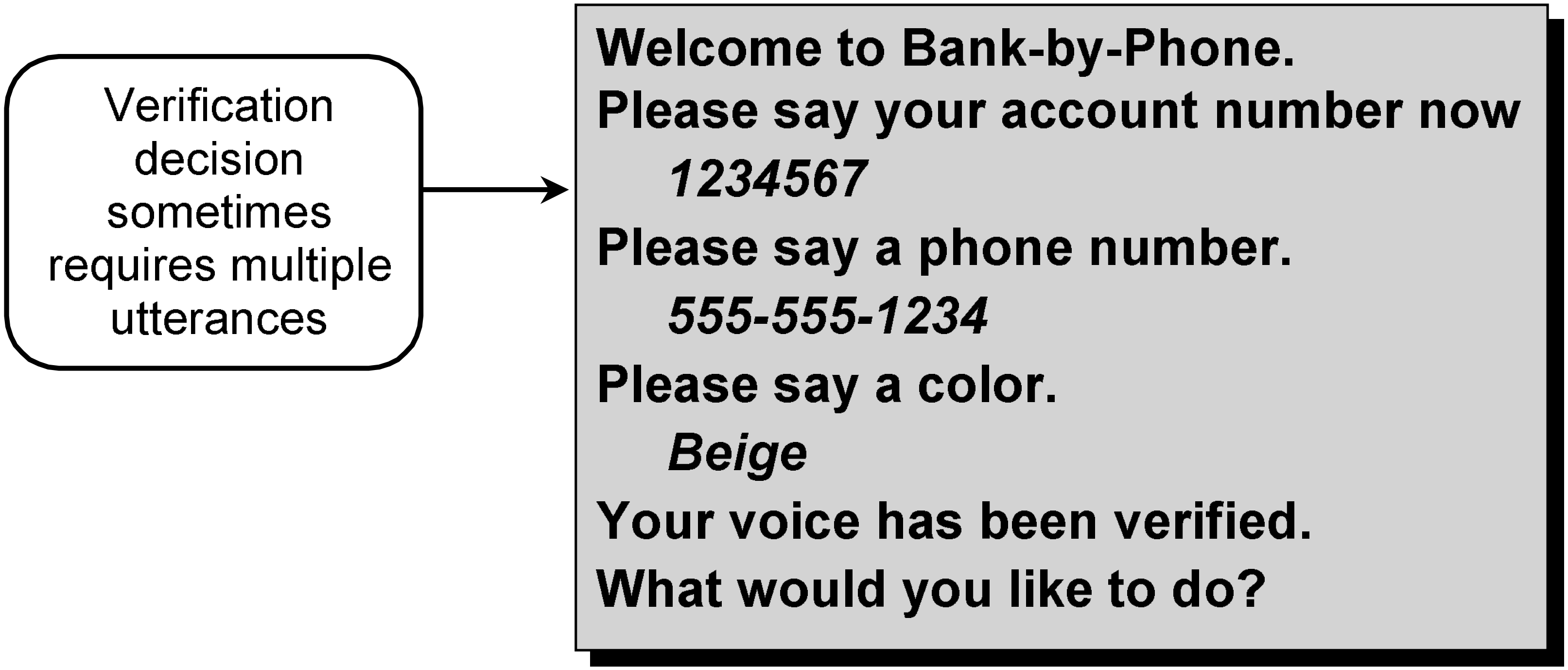
What would you like to do?

- **Note that ASR result determines which voiceprint to check**

Functionality

Default: Multi-utterance verification (2)

- **Verification accuracy improves with more user utterances.**



Verification decision sometimes requires multiple utterances

Welcome to Bank-by-Phone.

Please say your account number now

1234567

Please say a phone number.

555-555-1234

Please say a color.

Beige

Your voice has been verified.

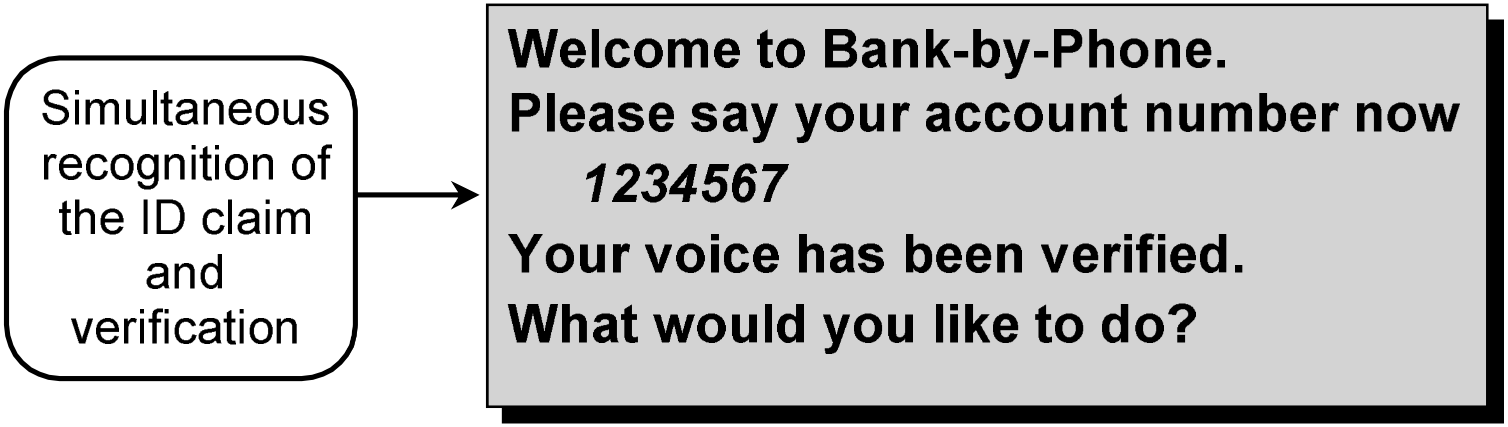
What would you like to do?

- **Note that ASR result determines which voiceprint to check**

Functionality

Simultaneous ID claim + verification

- **Permits single-step identification and authentication**



Simultaneous recognition of the ID claim and verification

Welcome to Bank-by-Phone.

Please say your account number now

1234567

Your voice has been verified.

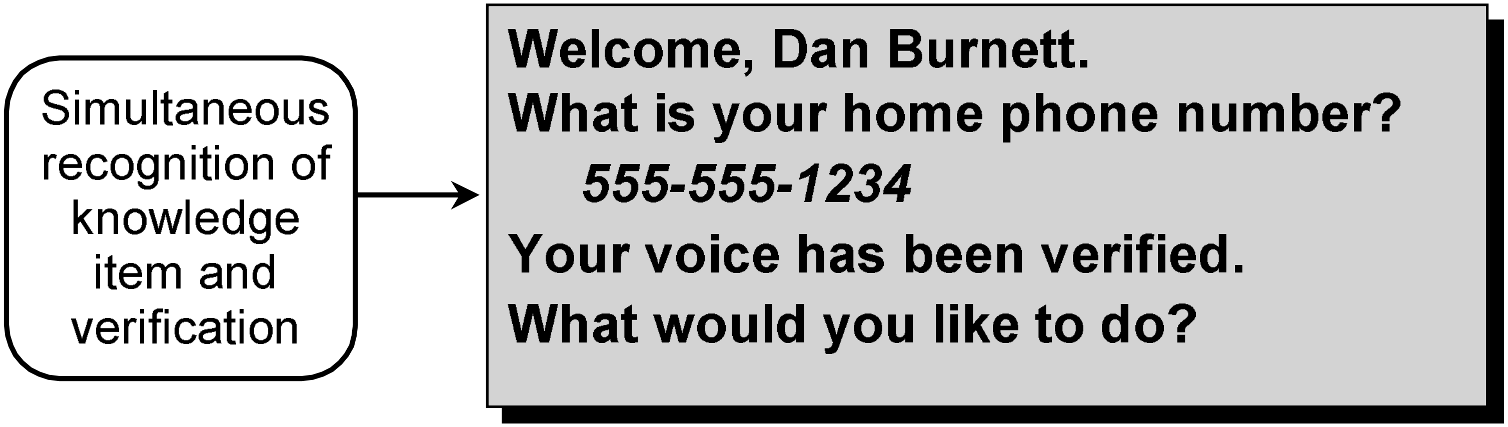
What would you like to do?

- **Note that ASR result determines which voiceprint to check**
- **Same audio used for recognition and verification so only one user utterance is required**

Functionality

Simultaneous knowledge and voice authentication

- Permits single-step but dual-mode (higher security) authentication



Simultaneous recognition of knowledge item and verification

Welcome, Dan Burnett.

What is your home phone number?

555-555-1234

Your voice has been verified.

What would you like to do?

- ID already available (through previous recognition, caller id, etc.)
- Note that ASR result verifies knowledge item
- Same audio used for recognition and verification so only one user utterance is required

Speaker Recognition
Outline

- **Overview of area**
  - Introduction/Terminology
  - General Theory
- **Needed application functionality**
  - Motivation: dialog design
  - Multi-utterance verification
  - Simultaneous identity claim and verification
  - Simultaneous knowledge and identity verification
- **Requirements**
  - Support for simultaneous ASR/verification/identification
  - Support integrated ASR/verification/identification
  - Support identified (named) resources

Requirements
Support simultaneous ASR/verification/ID

- **SRCP MUST enable simultaneous sending and control of caller audio stream to ASR, verification, and identification resources, because**
  - **Simultaneous ID claim and verification requires it**
  - **Simultaneous knowledge and voice authentication requires it**

  - **This is the <u>only</u> way to support rolling or dynamically-generated challenge phrases (e.g., “say 51723”)**

Requirements
Support integrated ASR/verification/ID

- **SRCP MUST enable sending and control of caller audio stream to an integrated ASR, verification, and identification resource, because**
  - Processing for these three technologies can reasonably be shared for better performance
    - Even small time delays in spoken dialogs are fatal (different from most graphical interfaces)
  - Technologies used together so often that it would be wasteful to
    - Require the use of multiple identical audio streams when processed by same resource
    - Require the use of many synchronization messages for activities often occurring simultaneously (comparison would be sync messages between ASR/TTS servers to kill prompt on barge-in but on grander scale)

Requirements
Support identified (named) resources

- **SRCP MUST provide ID for each verification resource and permit control of that resource by ID, because**
  - **Voiceprint format and contents are vendor-specific (once you select a resource you need to keep using it)**
  - **Resource must maintain state to handle multi-utterance verification**

Closing
More info

- The material from this presentation was largely taken from an ICASSP tutorial. For more details, see

L.P. Heck and D. Reynolds,
"Speaker Verification: From Research to Reality",
International Conference on Acoustics, Speech, and Signal Processing
Tutorial, Salt Lake City, Utah, May 2001.