Odyssey workshop 2001
Open discussion on algorithmic issues

Moderators:
Frederic BIMBOT & Douglas REYNOLDS

Contributors: ALL Odyssey workshop participants
Menu

1. Where can we expect IMPROVEMENTS in current state-of-the-art approaches?

2. What type of UNEXPLOITED sources of INFORMATION should we use, and how to do so?

3. Which are the most dramatic ROBUSTNESS issues that we are / will be facing and how can we cope with them?
Objective of the discussion

• Prepare a collective contribution to a Scientific Journal comprising:
  - A summary of today’s discussion
  - Several short (2-page) individual contributions defending a specific challenging issue

"Where are we, and where do we go from here?"

• Targets:
  - Within the scientific community: increase interest and excitement of researchers, students, ...
  - Outside the scientific community: underline margins and potential for future progress
Components of Verification System

- Features
- Models
- Score function
- Decision
Acoustic-based Systems

• **Features**
  - Cepstra and delta-cepstra, ...
  - $F_0$

• **Models**
  - Probabilistic
    - HMM, GMM, Gaussian, DTW, VQ, ...
  - Neural Net (Discriminative)
    - ANNN, MLP, etc...
  - SVM

• **Score functions**
  - Znorm, Hnorm, Tnorm, Htnorm, Posterior Probabilities, ...

• **Decisions**
  - Batch, Sequential, Fusion, ...
What info ISN’T being used currently?

- **Low-level (signal) features**
  - Speaker-specific signal components / representation
- **Suprasegmental features**
  - Patterns over time (acoustic, phonetic, …)
- **Prosodics**
  - F0 contours; pause patterns; rate of speech
- **Word usage**
  - Idiolectal, lexical, syntax
- **Non-linguistic**
  - Accents, laughs, idiosyncrasies
- **Behavioral**
  - Speaker-dependent characteristic response / interaction

How to use this info?
Robustness issues

• **Speaker itself**
  – Physiology, health, emotion, …

• **Type of device**
  – Phone (regular, mobile, internet / VOIP …)
  – Type of coding (variable rate coding, …)

• **Context of use / Environment**
  – Surrounding noises, background music
  – Other speakers
  – Old recordings

• **Lack of a priori knowledge**
  – Number of speakers
  – Type of channel

How are our algorithms going to resist these factors?
Open questions in current State-of-the-Art

• **Acoustic features**
  - How can we decouple speaker and channel effects?
  - How can we “customize” the features to the speaker?

• **Models**
  - Other models / frameworks
  - Where do SVM fit?
  - Are there ways to combine / embed components
    • Features and models
    • Models and score function

• **Score function / normalization**
  - What is a score function?
  - Do we understand “normalization”

• **Decision**
  - Is the Bayesian framework optimal in all circumstances?
  - Can we make “smarter” decisions?
What info ISN’T being used currently?

- **Low-level (signal) features**
  - Speaker-specific signal components / representation
- **Suprasegmental features**
  - Patterns over time (acoustic, phonetic, …)
- **Prosodics**
  - F0 contours; pause patterns; rate of speech
- **Word usage**
  - Idiolectal, lexical, syntax
- **Non-linguistic**
  - Accents, laughs, idiosyncrasies
- **Behavioral**
  - Speaker-dependent characteristic response / interaction

How to use this info?
Robustness issues

• **Speaker itself**
  – Physiology, health, emotion, …

• **Type of device**
  – Phone (regular, mobile, internet / VOIP …)
  – Type of coding (variable rate coding,…)

• **Context of use / Environment**
  – Surrounding noises, background music
  – Other speakers
  – Old recordings

• **Lack of a priori knowledge**
  – Number of speakers
  – Type of channel

How are our algorithms going to resist these factors?
Tasks

What are the Tasks in speaker recognition

- Close set ID
- Open-set ID
- Verification
- Segmentation / tracking
- Numbering / clustering
- Matching (adaptation speech recognition)