Time-Varying Speaker Recognition

An Introduction

Wang Linlin

Nov. 26th, 2012

Speaker recognition

- Voiceprint recognition
- One kind of biometric authentication technology
  - by using speaker-specific information contained in speech waves
  - “non-contact, non-intrusive and easy to use”
  - ranked first by consumer preference among biometric measures according to a Unisys survey

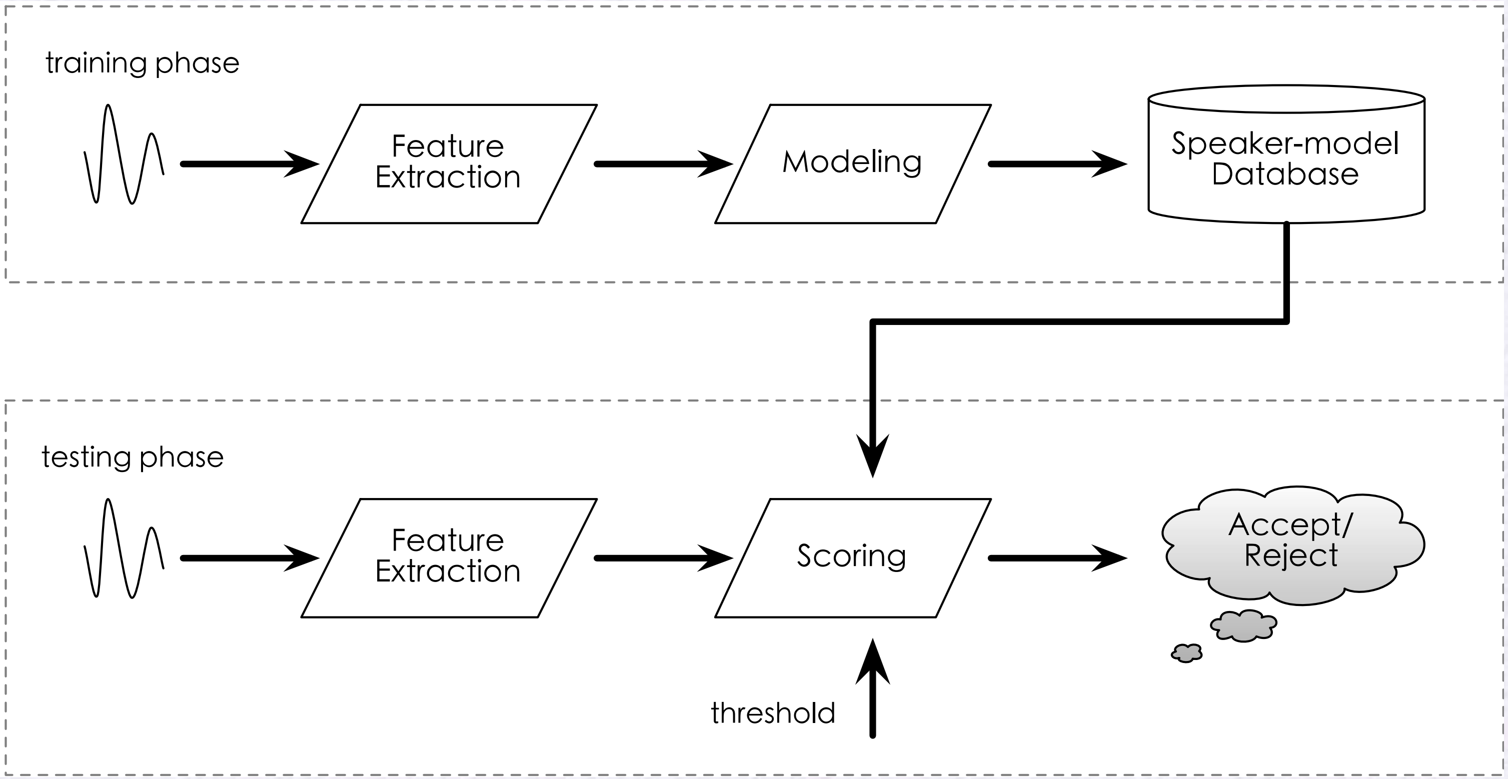
Practical Applications

- Commercial
  - ABN AMRO & Voice Vault
  - NAB & Telstra Salmat VeCommerce
  - CCB & d-Ear
- Public service
  - Wellpoint
- Public & national security

Challenges

- Common ones in speech-related technologies
  - Poor-quality voice samples
  - Background noise
  - Channel mismatch
- Specific ones in speaker recognition technology
  - Short utterance
  - Within-speaker variability
    - Speaking style, emotion, physical status, changes over time…
  - ……

Framework



Framework of a classic speaker verification system

Time-varying Issue

- “Does the voice of an adult change significantly with time? If so, how?” [Kersta 1962]
- “How to deal with long-term variability in people’s voice?” [Furui 1997]
- “Voice changes over time, either in the short-term, the medium-term, or in the long-term.” [Bonastre *et al*. 2003]

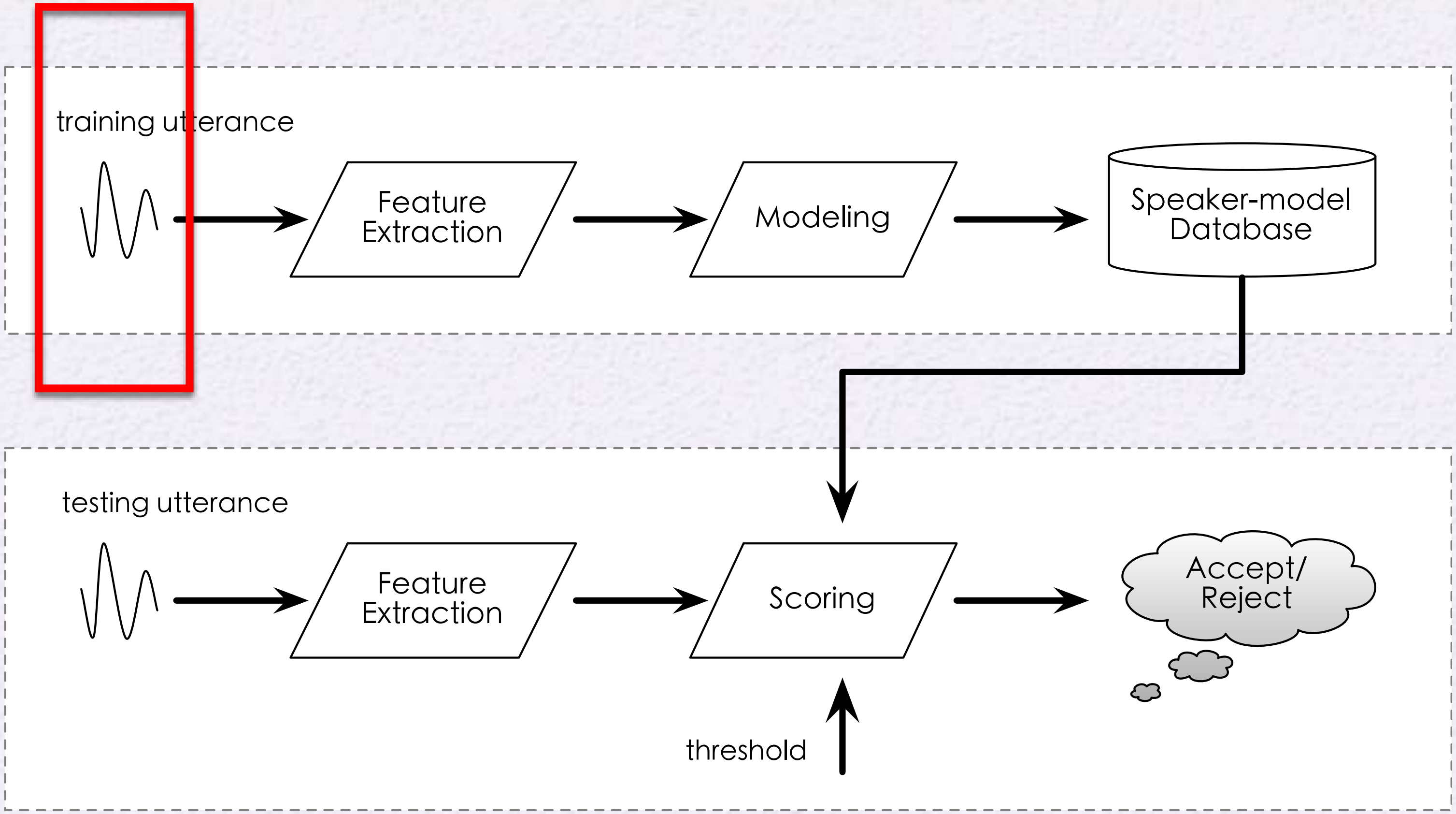
Time-varying Issue

- Performance degradation
  - "The longer the separation between the training and testing recordings, the worse the performance." [Soong *et al.* 1985]
  - A significant loss in accuracy between two sessions separated by 3 months
    - 4~5% in EER [Kato & Shimizu 2003]
    - Ageing was considered to be the cause [Hebert 2008].
  - A voiceprint access control system in CCNT lab
    - 69.02% to 74.19% [Shan & *Yang* 2005]

Time-varying Issue

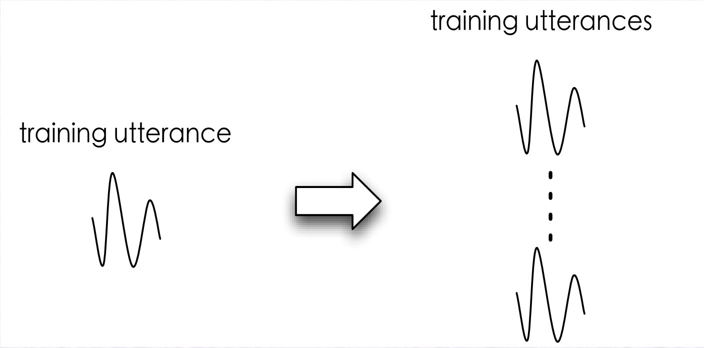
- A generally acknowledged phenomenon
  - Speaker recognition performance degrades with time varying.
  - "Mysterious factors" [Kenny et al. 2007]
- How to deal with this issue?

Methods


training utterance
Feature Extraction
Modeling
Speaker-model Database
testing utterance
Feature Extraction
Scoring
Accept/ Reject
threshold

Structural Training

- More training data lead to more representative models
- Several researchers resorted to several training sessions over a long period of time to help coping with the long-term variability of speech.
  - [Bimbot *et al.* 2004]
  - [Soong *et al.* 1985]



Reference Set

- The best speaker recognition result was obtained when 5 sessions successively separated by at least 1 week were used to define the reference (training) set. [Markel and Davis 1979]

PERCENT OF SPEAKERS CORRECTLY IDENTIFIED AS A FUNCTION OF THE NUMBER OF REFERENCE SESSIONS

SESSIONS			L_V			
NO.	REF.	TEST	30	100	300	1000
2	1-2	3-4	50.36	64.34	71.18	—
2	3-4	1-2	53.45	67.95	75.31	—
3	1-3	4-6	54.29	70.03	79.12	80.58
3	4-6	1-3	57.04	72.69	82.14	89.30
4	1-4	5-8	59.91	76.41	86.73	92.85
4	5-8	1-4	59.26	74.62	83.45	86.34
5	1-5	6-10	61.20	78.65	88.20	93.34
5	6-10	1-5	59.87	75.48	85.27	89.77

Data Augmentation

- When a positive identification of the candidate speaker is made, extra data is appended to the original enrollment data to provide a more universal enrollment model for the candidate.[Beigi 2009] [Beigi

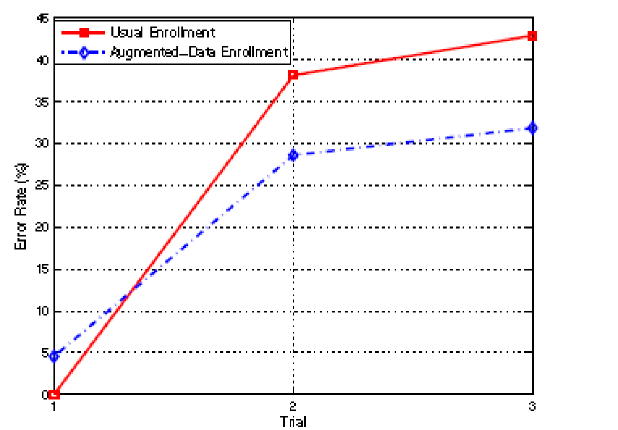
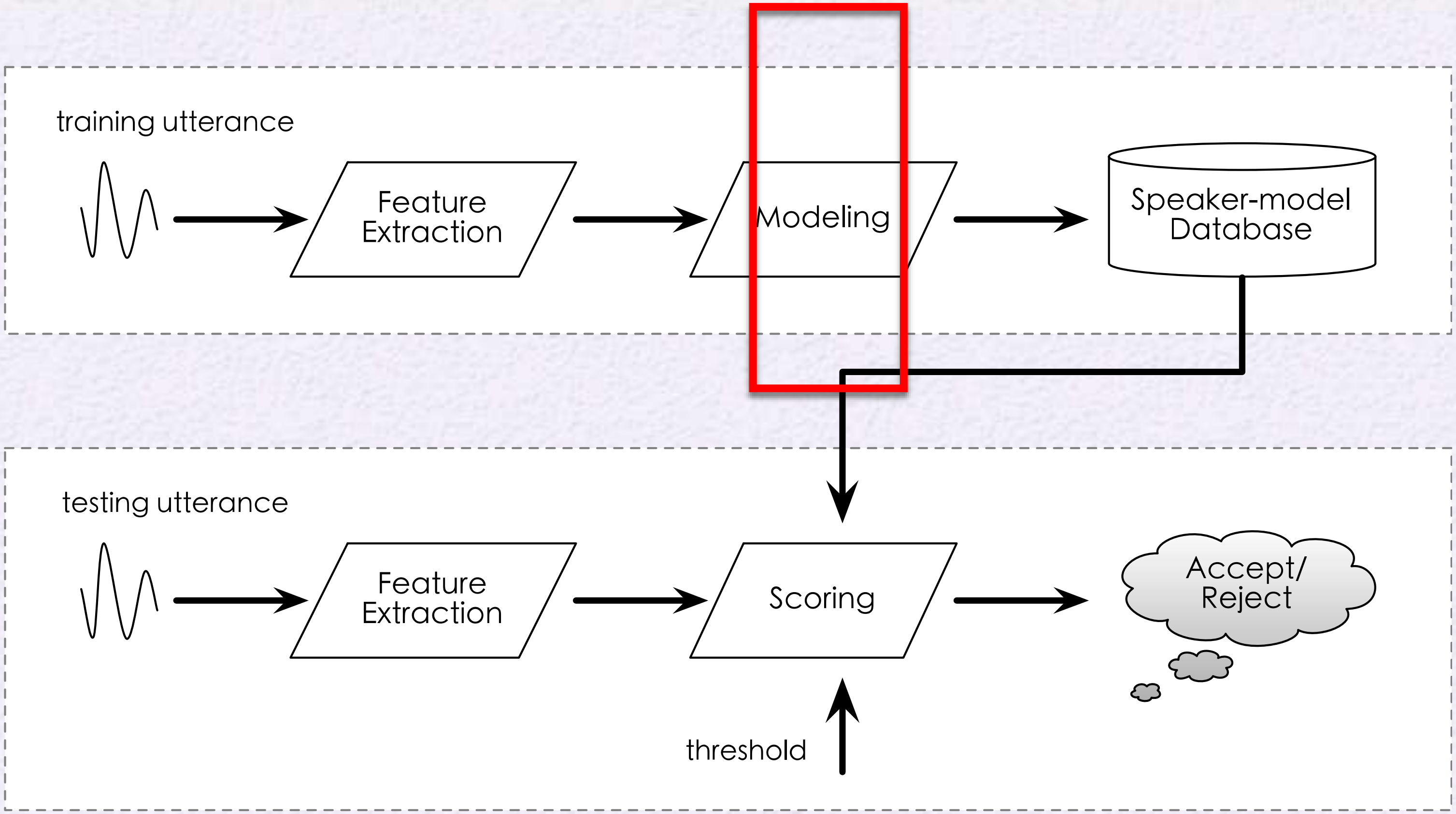


Fig. 3. Identification Time Lapse – Augmented-Data Enrollment

Methods


training utterance
Feature Extraction
Modeling
Speaker-model Database
testing utterance
Feature Extraction
Scoring
Accept/ Reject
threshold

Model Adaptation

- To use MAP adaptation to adapt from the original model to a new model considering new data at hand. [Beigi 2009] [Beigi 2010]

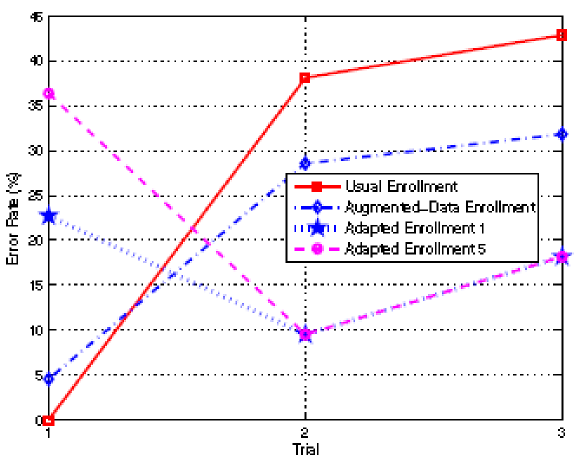
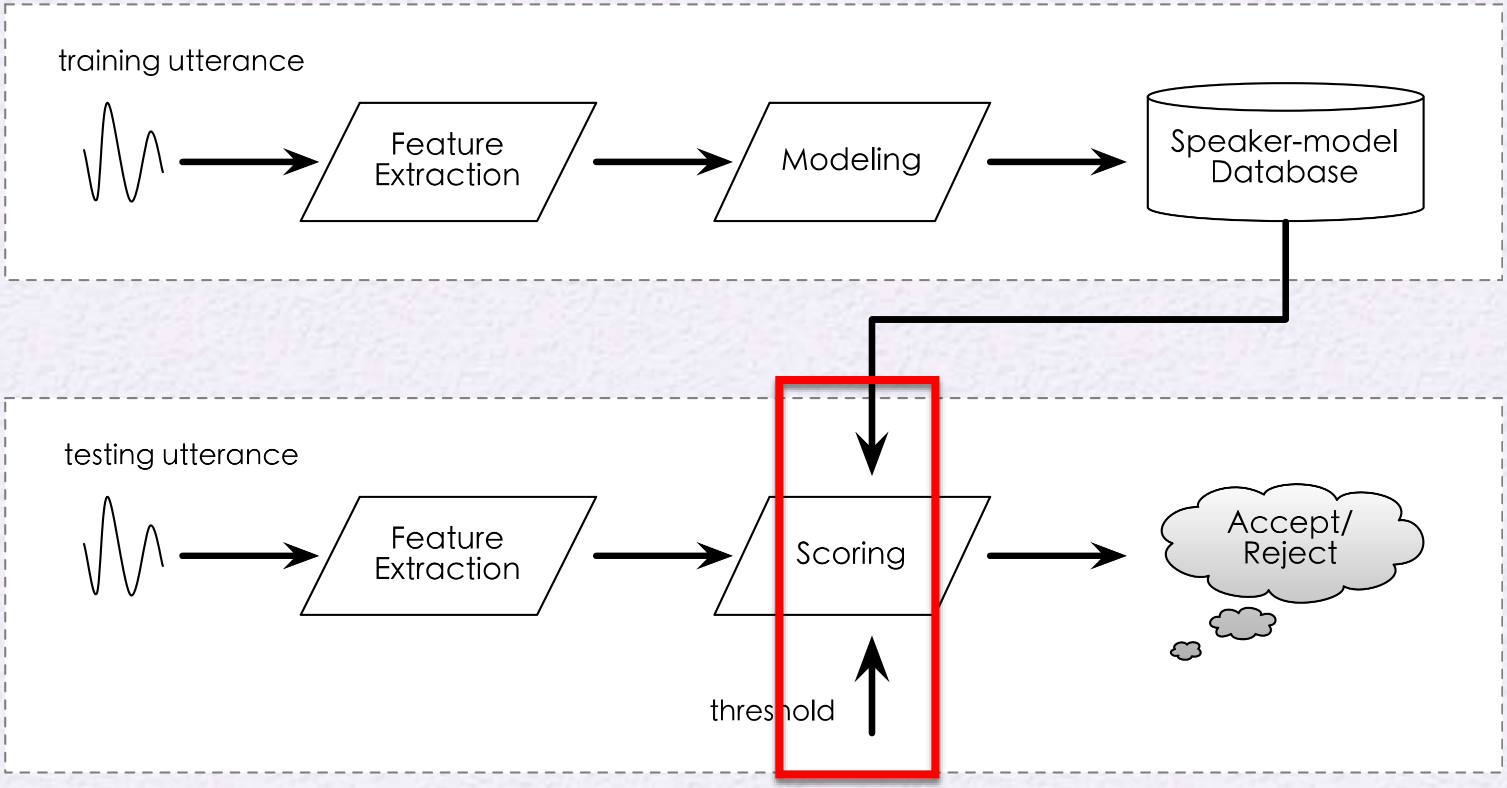


Fig. 5. Identification Time Lapse

Model Adaptation

- To use MLLR-based speaker-adaptation technique to reduce the effects of model aging.
  - [Lamel & Gauvin 2000]
  - EER on the last two sessions is reduced to 1.7% from 2.5%, after adapting the speaker models on data from the intervening session.

Methods


training utterance
Feature Extraction
Modeling
Speaker-model Database
testing utterance
Feature Extraction
Scoring
Accept/ Reject
threshold

Threshold Decision

- Verification scores of genuine speakers decreased progressively as the time span between training and testing increases, while impostor scores were less affected. [Kelly & Harte 2011] [Kelly *et al.* 2012]

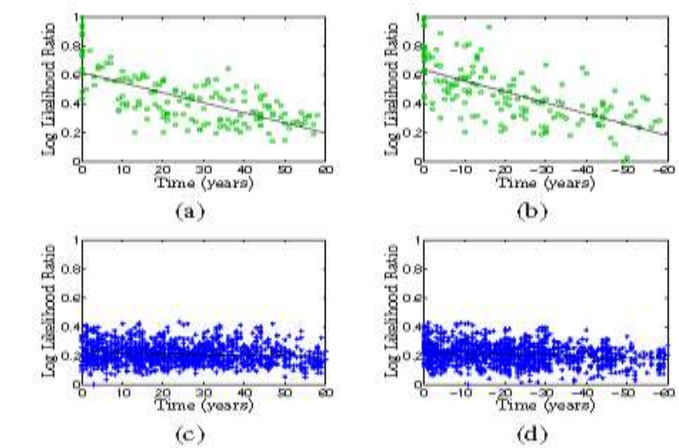


Figure 1. LLR scores for all 18 speakers: (a) genuine speakers, forwards, (b) genuine speakers, backwards, (c) imposters, forwards, (d) imposters, backwards

Threshold Decision

- A stacked classifier method of introducing an ageing-dependent decision boundary was applied, significantly improving long-term verification accuracy. [Kelly & Harte 2011] [Kelly *et al.* 2012]

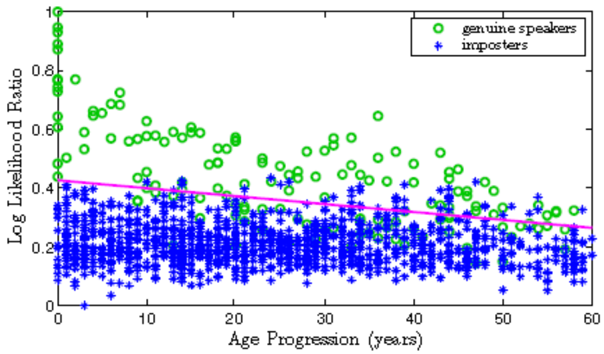
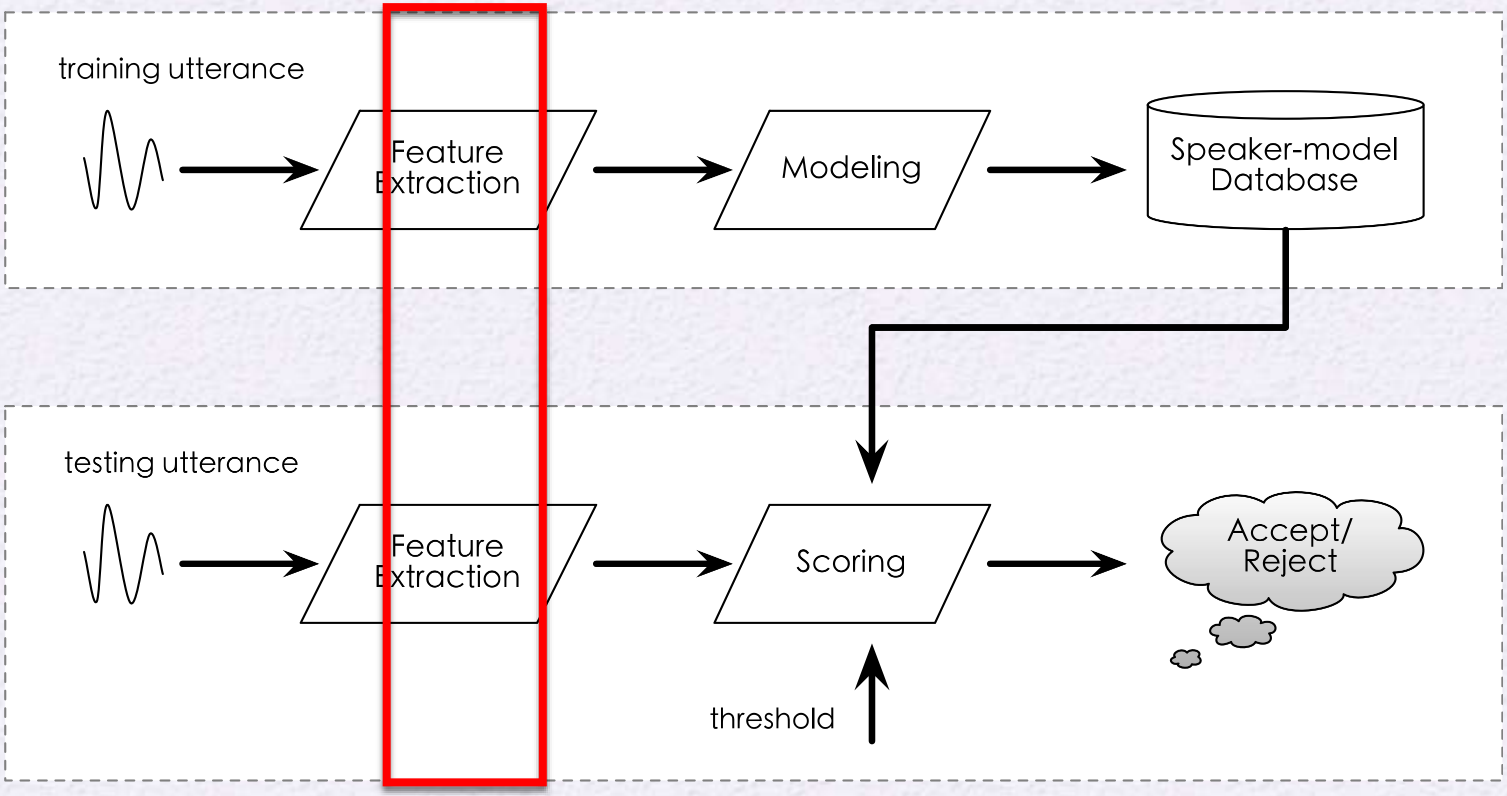


Figure 3. A global ageing-dependent decision boundary trained from 17 speakers' LLR scores and age progression information

Age Progression (years)	5	10	20	40	60
Forwards					
Fixed Threshold	10.8	15.9	26.5	32.2	36.1
Ageing-dependent	7.3	9.2	10.4	17.2	17.5
Backwards					
Fixed Threshold	13.7	18.1	18.9	24.8	29.1
Ageing-dependent	10.2	11.4	12.3	17	21.9

Table 1. Average HTER for all 18 speakers in the Speaker Ageing Database across different ranges of age progression.

Methods



training utterance
Feature Extraction
Modeling
Speaker-model Database
testing utterance
Feature Extraction
Scoring
Accept/ Reject
threshold

Feature

- The CORE problem in pattern recognition [Huang *et al.* 2001]
- An IDEAL feature for speaker recognition [Kinnunen & Li 2010] [Rose 2002] [Wolf 1972]
  - Have large between-speaker variability and small within-speaker variability
  - Not be affected by long-term variations in voice

More Stable Features

- Fundamental frequency generally fluctuates randomly across time-varying sessions. [Chen & Yang 2010] [Lu 2008]
- SMFCC [Lu 2008]
  - Smooth the amplitude spectrum and calculate the spectral envelope
  - Gender-dependent performance
    - It works better in female case, and not so good in male case.

An Analysis

- An ideal case
  - Users of speaker recognition systems log-in from time to time, to update their models
  - Advantages
    - Utterances from the genuine speaker
    - "Up-to-date" models
  - Disadvantages
    - User-unfriendly, extra burden

An Analysis

- Structural Training & Model Adaptation
  - More training data and extra adaptation data
  - Advantages
    - No extra burden for users
  - Disadvantages
    - Higher requirements on systems
      - A longer registration process
      - “Threshold” of utterances from the genuine speaker
    - Blind update
      - Nothing to do with the NATURE of time-varying

An Analysis

- Ageing-dependent decision boundary and SMFCC
  - Solutions regarding the trends how fundamental frequency and verification score change over time
  - Disadvantages
    - Both have their own restrictions
  - Advantages
    - This kind of “targeted” attempts should be a natural research direction in time-varying speaker recognition

Data Matters

- Trends are obtained from careful data analysis.
- A proper longitudinal voiceprint database is needed for time-varying research in speaker recognition, which will be elaborated in my second presentation.

References

- L.G. Kersta, "Voiceprint recognition", *Nature*, no. 4861, pp. 1253-1257, December 1962.
- S. Furui, "Recent advances in speaker recognition", *Pattern Recognition Letters*, vol. 18, iss. 9, pp. 859-872, September 1997.
- J. Bonastre, F. Bimbot, L. Boe, *et al.*, "Person authentication by voice: a need for caution", *Proc. of Eurospeech 2003*, pp. 33-36, Geneva, 2003.
- F. Soong, A. E. Rosenberg, L. R. Rabiner, *et al.*, "A vector quantization approach to speaker recognition", *Proc. of ICASSP 1985*, vol.10, pp. 387-390, Florida, 1985.
- T. Kato, and T. Shimizu, "Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence preserving connected digit patterns", *Proc. of ICASSP 2003*, Hong Kong, 2003.
- M. Hebert, "Text-dependent speaker recognition", *Springer Handbook of Speech Processing*, Springer-Verlag: Berlin, 2008.
- 单振宇，杨莹春，“声纹打卡系统”，*第八届全国人机语音通讯学术会议*，pp. 565-568，2005.
- P. Kenny *et al.*, "Joint factor analysis versus eigenchannels in speaker recognition", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, iss. 4, pp. 1435-1447, May 2007.
- F. Bimbot *et al.*, "A tutorial on text-independent speaker verification", *EURASIP Journal on Applied Signal Processing*, vol. 2004, iss. 1, pp. 430-451, January 2004.
- J. Markel and S. Davis, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 1, pp. 74-82, February 1979.
- H. Beigi, "Fundamentals of speaker recognition", New York Springer, 2010.
- H. Beigi, "Effects of time lapse on speaker recognition results", *Proc. of 16th International Conference on Digital Signal Processing*, pp. 1-6, 2009.
- L. Lamel and J. Gauvin, "Speaker verification over the telephone", *Speech Communication*, vol. 2000, iss. 31, pp. 141-154, 2000.
- F. Kelly and N. Harte, "Effects of long-term ageing on speaker verification", *Biometrics and ID Management*, Volume 6583 of Lecture Notes in Computer Science, pp. 113-124, Springer Berlin/Heidelberg, 2011.
- F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification with long-term ageing data", *Proc. of 5th IAPR International Conference on Biometrics*, New Delhi, 2012.
- X. Huang, A. Acero, and H. Hon, "Spoken language processing: a guide to theory, algorithm and system development", pp. 419-426, Prentice Hall, New Jersey, 2001.
- T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors", *Speech Communication*, vol. 2010, iss. 52, pp. 12-40, 2010.
- P. Rose, "Forensic speaker identification", Taylor & Francis London, 2002.
- J. Wolf, "Efficient acoustic parameters for speaker recognition", *Journal of Acoustic Society of America*, vol. 51, no. 6, pp. 2044-2056, 1972.
- 陈文翔，杨莹春，“声纹漂移现象初探”，第九届中国语音学学术会议，2010.
- 陆伟，“基于缺失特征的文本无关说话人识别鲁棒性研究”，博士学位论文，中国科技大学，2008.

Thanks