



THE UNIVERSITY  
of EDINBURGH



UNIVERSITY OF  
EASTERN FINLAND

# **Spoofing and anti-spoofing: a shared view of speaker verification, speech synthesis and voice conversion**

**APSIPA ASC tutorial, 16<sup>th</sup> Dec. 2015**

Zhizheng Wu, University of Edinburgh, UK

Tomi Kinnunen, University of Eastern Finland, Finland

Nicholas Evans, EURECOM, France

Junichi Yamagishi, University of Edinburgh, UK & National Institute of Informatics, Japan



# Presenters

Zhizheng Wu

---

Univ. of Edinburgh, UK



Tomi H. Kinnunen

---

UEF, Finland



Nicholas Evans

---

EURECOM, France



Junichi Yamagishi

---

Univ. of Edinburgh, UK  
NII, Japan



# Presentation material

<http://www.spoofingchallenge.org/apsipa/>



# Outline

## Part 1

1. Introduction
2. Speaker verification
3. Speech synthesis
4. Voice conversion
5. Q&A

## Part 2

6. Spoofing
7. Countermeasures
8. ASVspoof 2015
9. Future
10. Q&A



# 1. Introduction

biometrics

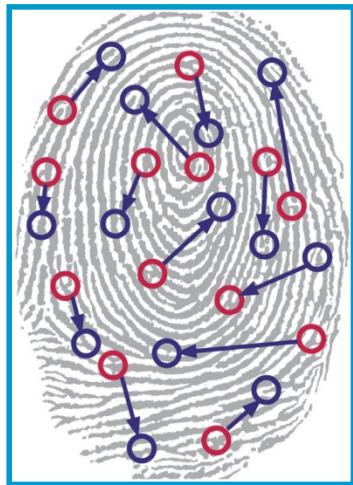
assessment

vulnerabilities

spoofing

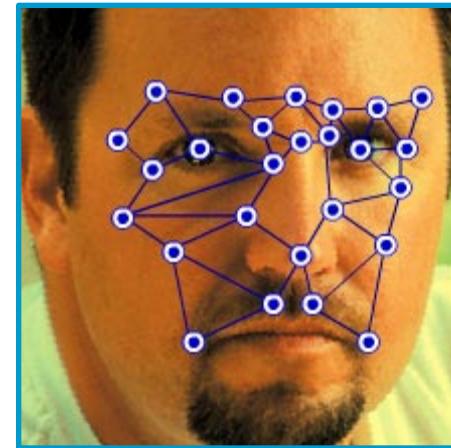
automatic speaker verificaiton

# Static / physiological modalities



Finger-print

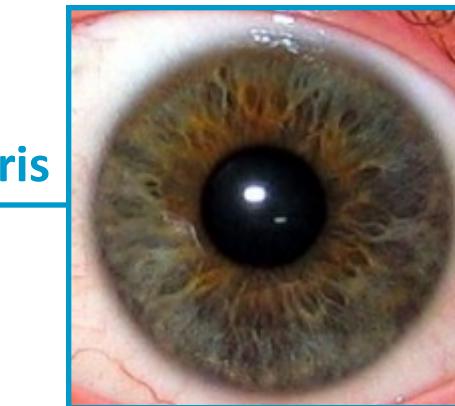
[biometricupdate.com](http://biometricupdate.com)



Face

[questbiometrics.com](http://questbiometrics.com)

- ICAO (International civil aviation authority) biometrics



Iris

[wikipedia.org](http://wikipedia.org) / Michael Reeve

# Dynamic / behavioural biometrics



Gait

source unknown

EEG/ECG



starlab.es



Signature

atvs.ii.uam.es

Speech



source unknown

# Applications

authentication



source unknown

surveillance



source unknown

differences in 'user' cooperation,  
but generally a common assessment methodology

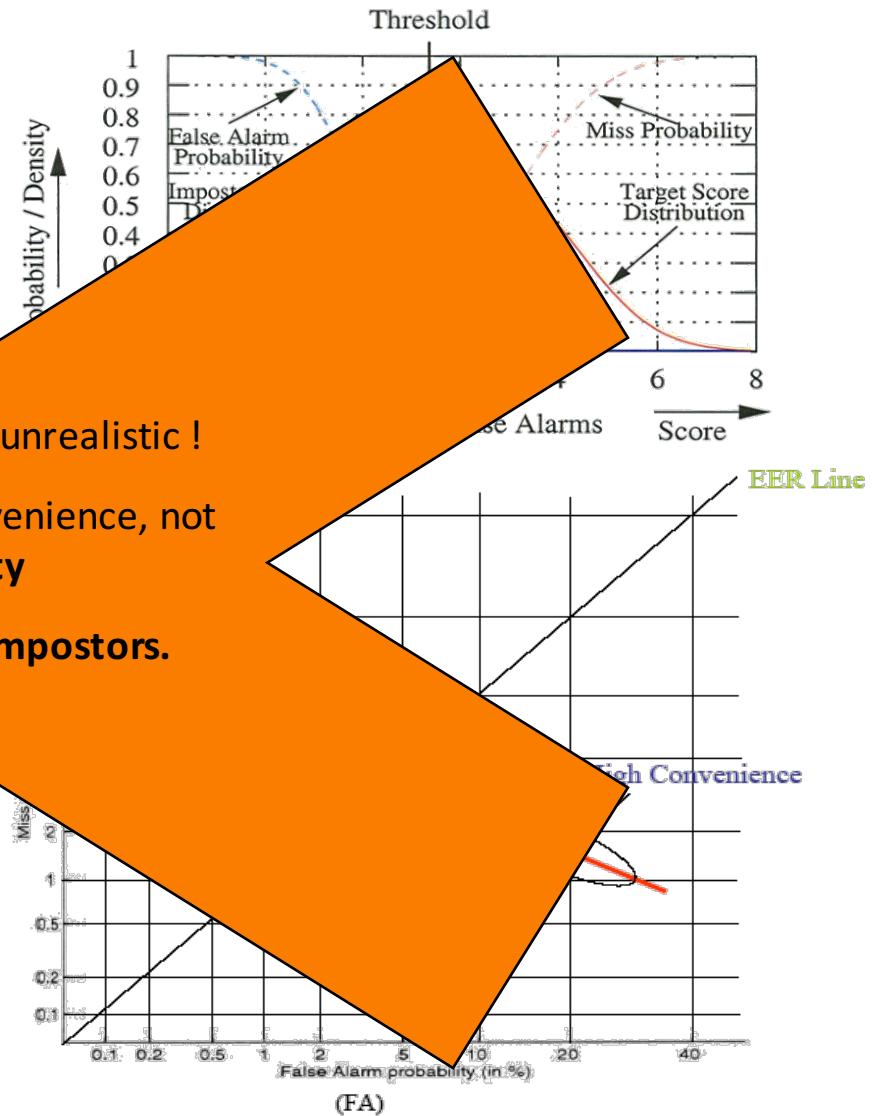
# Assessment



Score  
↓  
Threshold

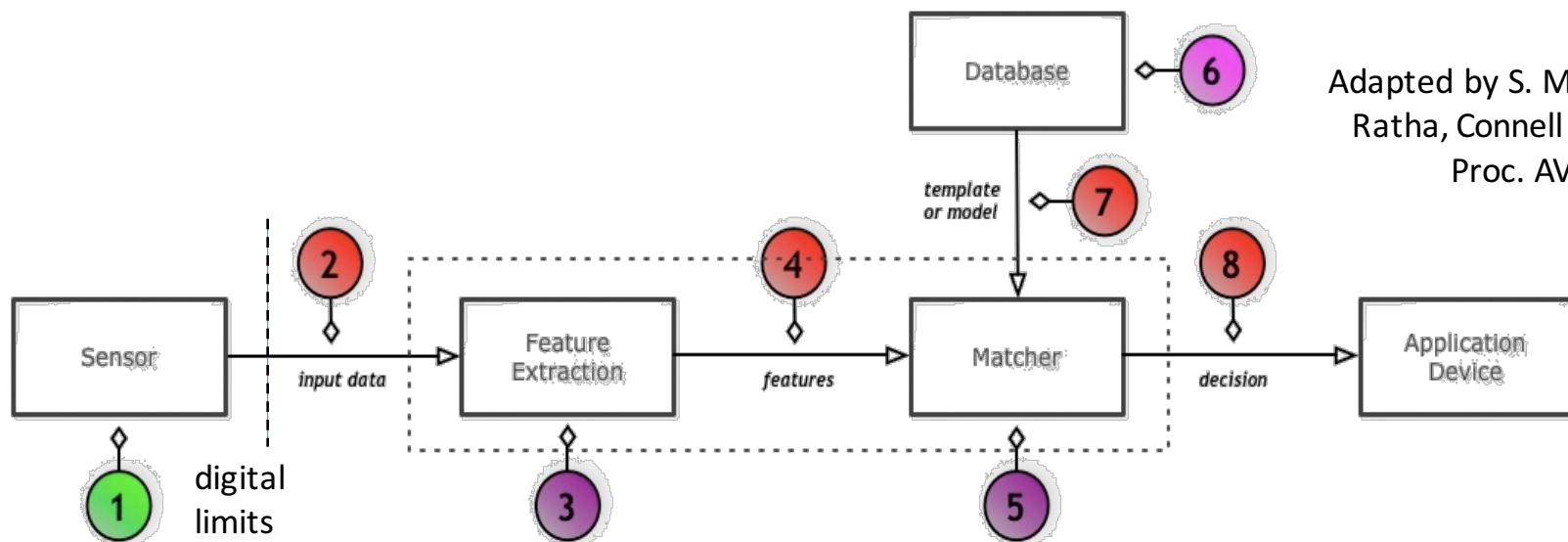
... it's potentially unrealistic !  
usability and convenience, not  
security  
... zero-effort impostors.

		Reject
Target	Correct verification	False Rejection
Impostor	False Alarm	Correct verification



# Biometric system vulnerabilities

- direct: prior to the digital limits (1)
- indirect: system intruders, hackers (2-8)



Adapted by S. Marcel from  
Ratha, Connell and Bolle,  
Proc. AVBPA, 2001

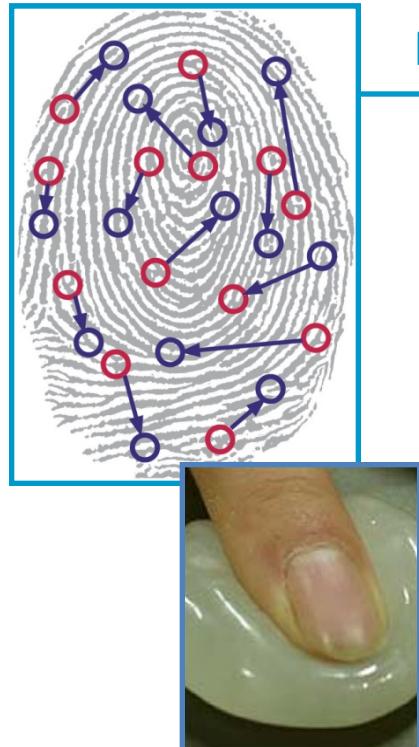
# Direct attacks: subversion / subterfuge

- concerted effort to deceive
- surveillance
  - evasion
  - obfuscation
  - provoke FRs
- authentication
  - **spoofing**
  - provoke FAs

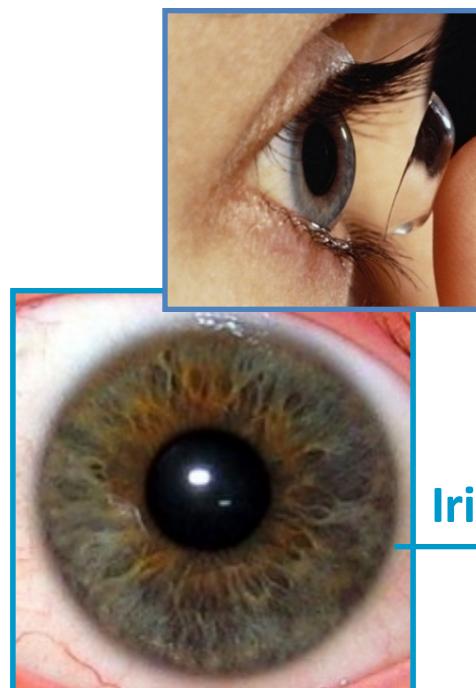


source unknown

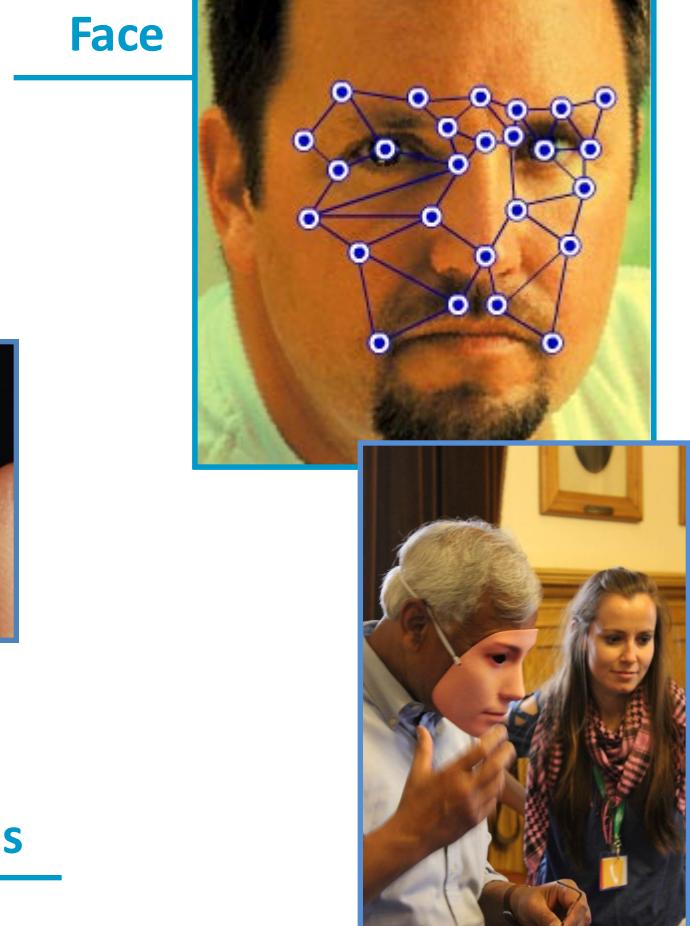
# Spoofing



Finger-print



Iris



Face

# Spoofing in the wild



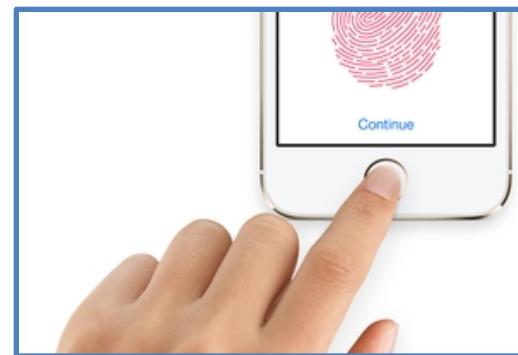
[biometricupdate.com](http://biometricupdate.com)

fingerprint  
recognition

face recognition  
(by a human)



EPA

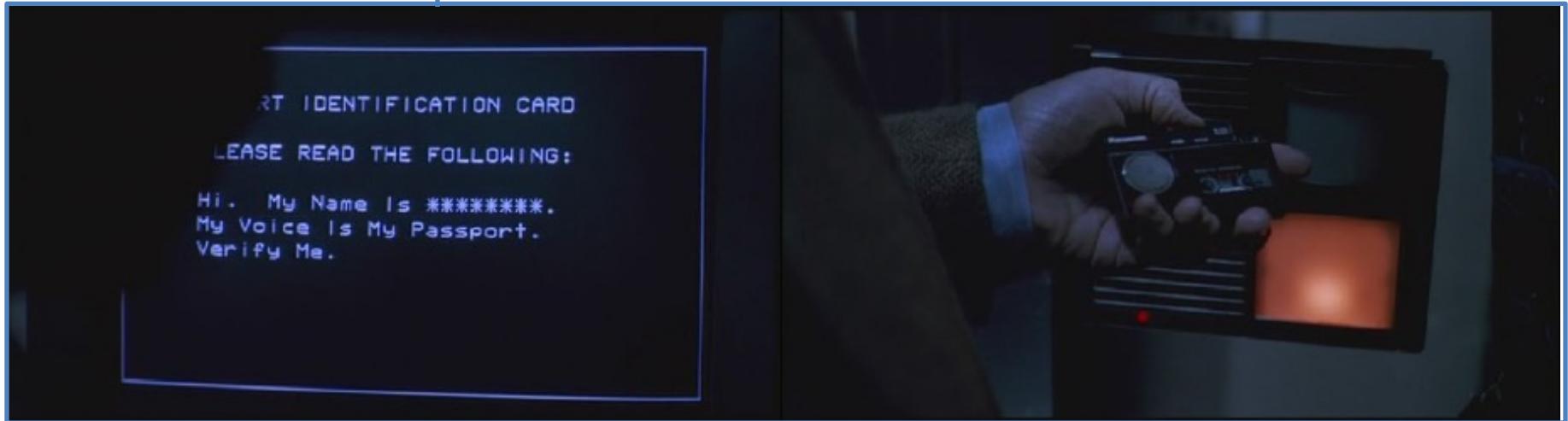


[planetbiometrics.com](http://planetbiometrics.com)

fingerprint  
recognition

# Speaker verification spoofing

replay spoofing – Sneakers 1992



Universal Pictures

# Speaker verification spoofing

- unattended, distributed scenarios
  - no human supervision
- approaches to spoofing
  - impersonation
  - replay
  - voice conversion
  - speech synthesis



How effective are they?



Can we detect them?

# The threat

- some spoofing attacks we **do** know about
- how many more do we **not** know about ?
- what is the threat / cost / damage / **risk**?
- what are we doing about it?

# ... towards ASVspoof

- standards
  - ICAO modalities only
  - datasets and evaluations
    - LivDET (fingerprint & iris), ICB (face)
- speaker verification
  - special session at Interspeech, 2013
  - IEEE SLTC newsletter article, 2013
  - **ASVspoof 2015**
- joint view of **verification**, SS, VC and spoofing

# Outline

## Part 1

1. Introduction
2. Speaker verification
3. Speech synthesis
4. Voice conversion
5. Q&A

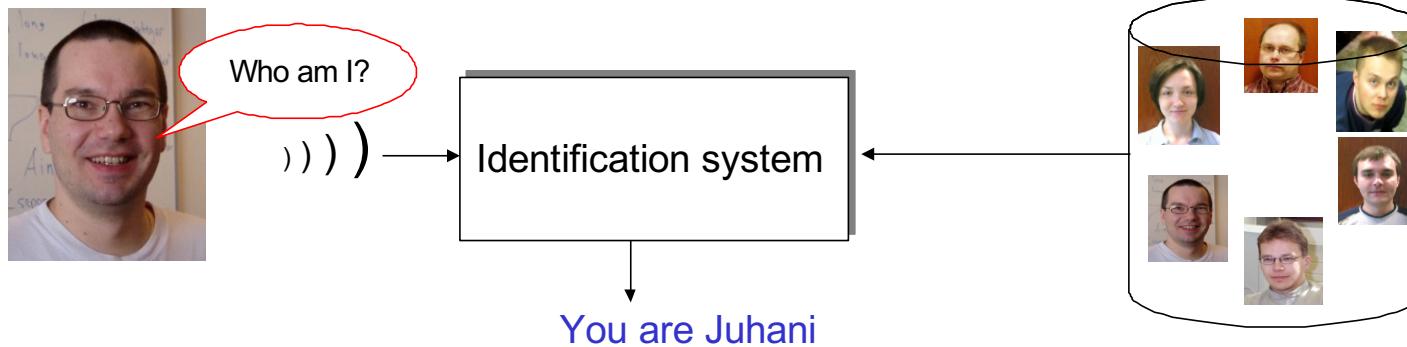
## Part 2

6. Spoofing
7. Countermeasures
8. ASVspoof 2015
9. Future
10. Q&A



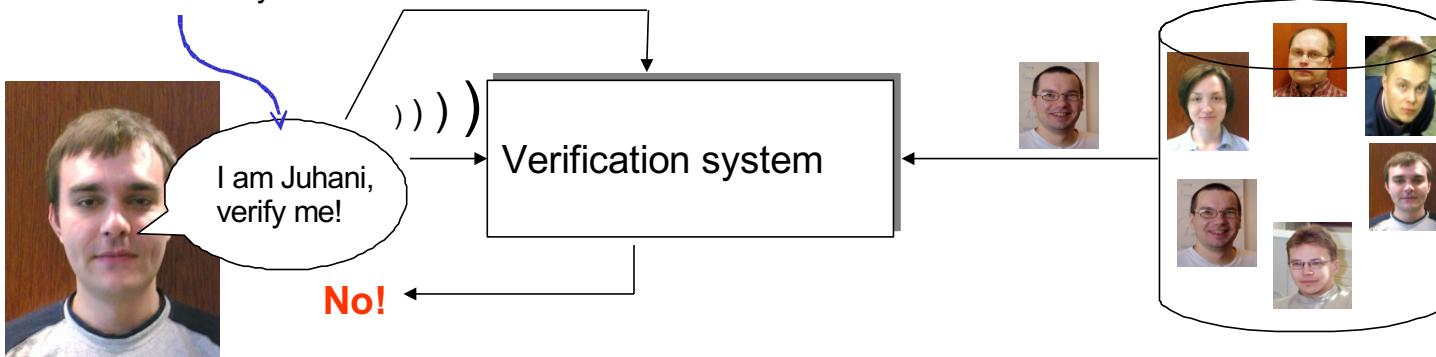
# Automatic Speaker Verification (ASV)

**Identification:** “Whom this voice belongs to?”



**Verification:** “Is this Juhani's voice?”

Claimed identity



Database of known  
speakers

# Application areas of ASV

- 1. User authentication – replacing passwords**
  - 2013: "Voice login" in Baidu-Lenovo phone
  - 2015: Similar efforts by Google using "ok google"
  - Call centers, banks
- 2. Forensics – voice evidence in telep. calls**
  - Shooting of Trayvor Martin (FL, US)
- 3. Surveillance / search / indexing**
  - Indexing multimedia archives
  - Intelligence, anti-terrorism

# The two operation modes

1. **Text-dependent (TD)** : enrolment and verification utterances share (at least partially) same content
2. **Text-independent (TI)** : arbitrary text in both enrolment and verification (even different language)

	Text-dependent	Text-independent
Authentication	X	X
Forensics		X
Surveillance/search		X

# Same or Different Speaker ?

**Speaker pair 1**



Different speaker

**Speaker pair 2**



Same speaker

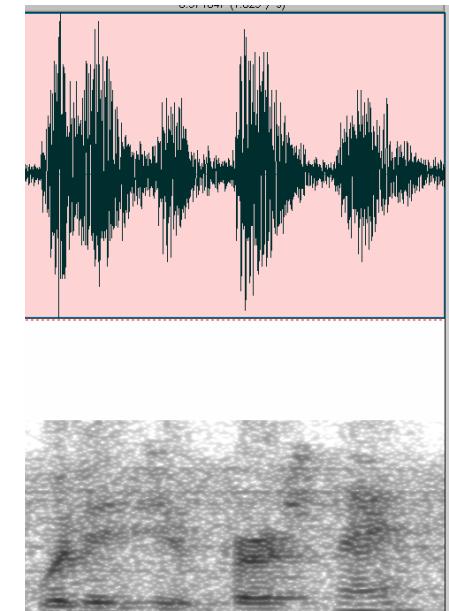
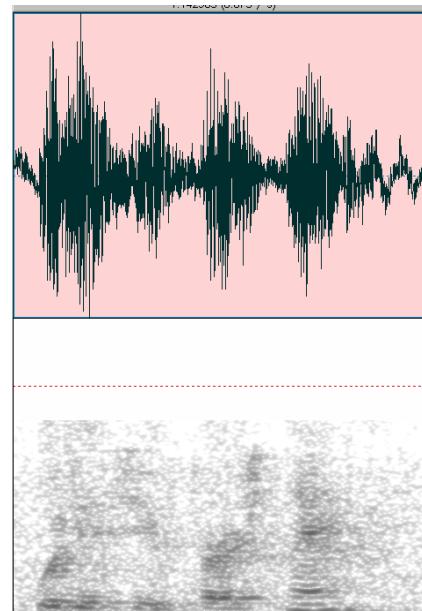
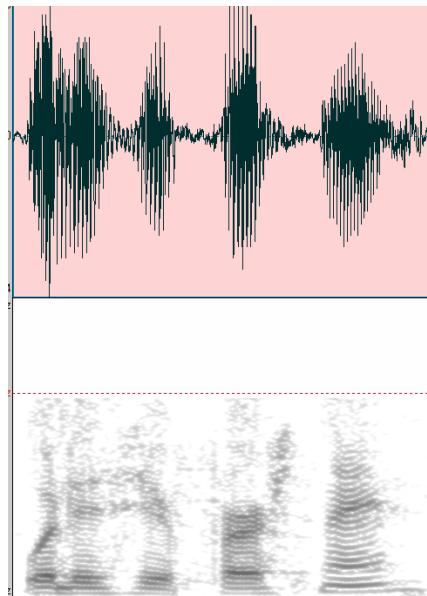
**Speaker pair 3**



Same speaker

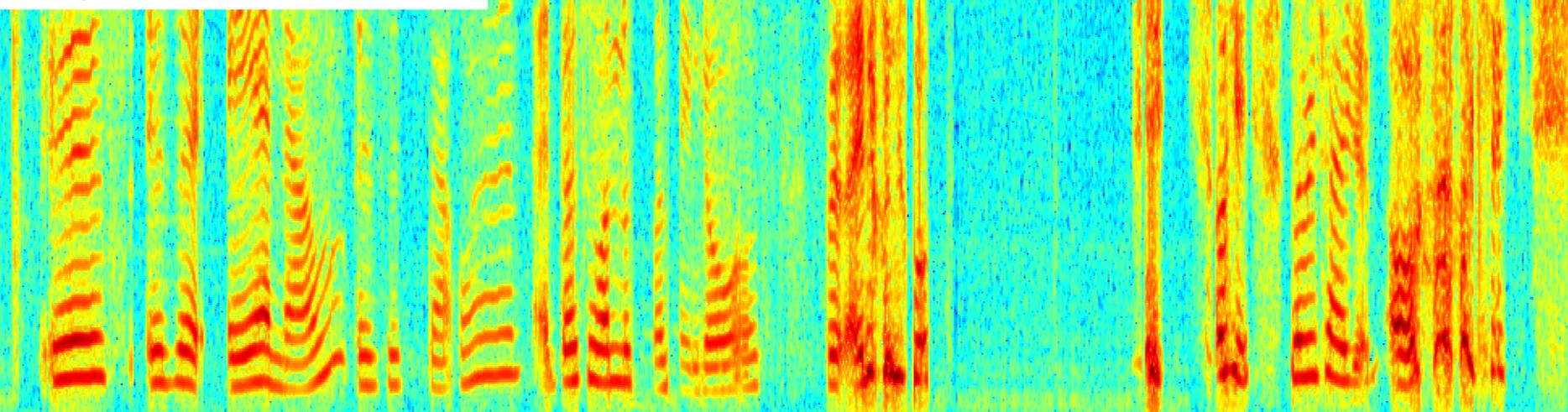
# Challenge: channel variation

The same source speech seen through three different channels

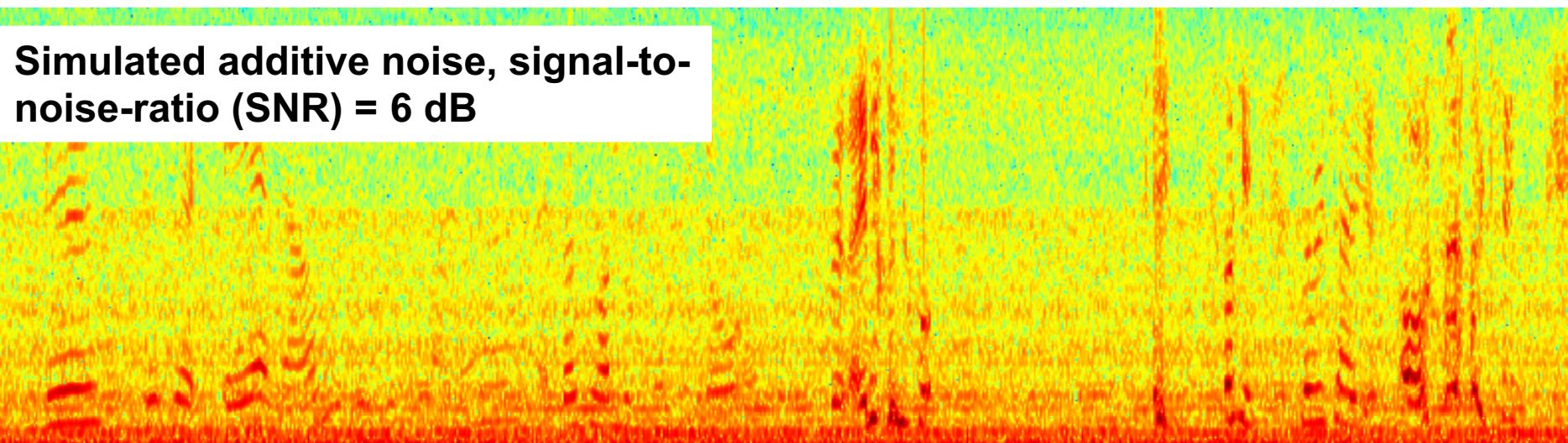


# Challenge: additive noise

Original utterance

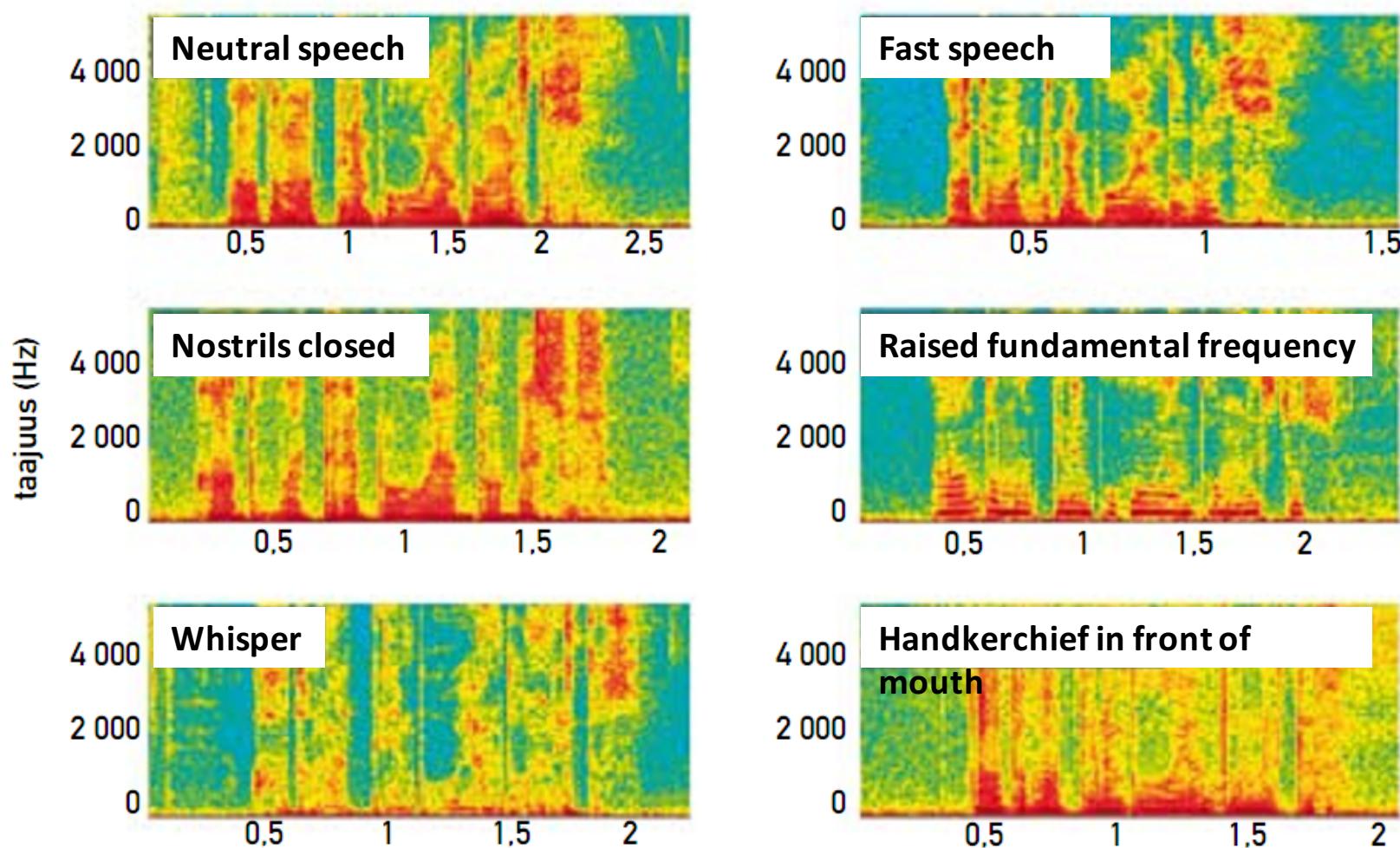


Simulated additive noise, signal-to-noise-ratio (SNR) = 6 dB



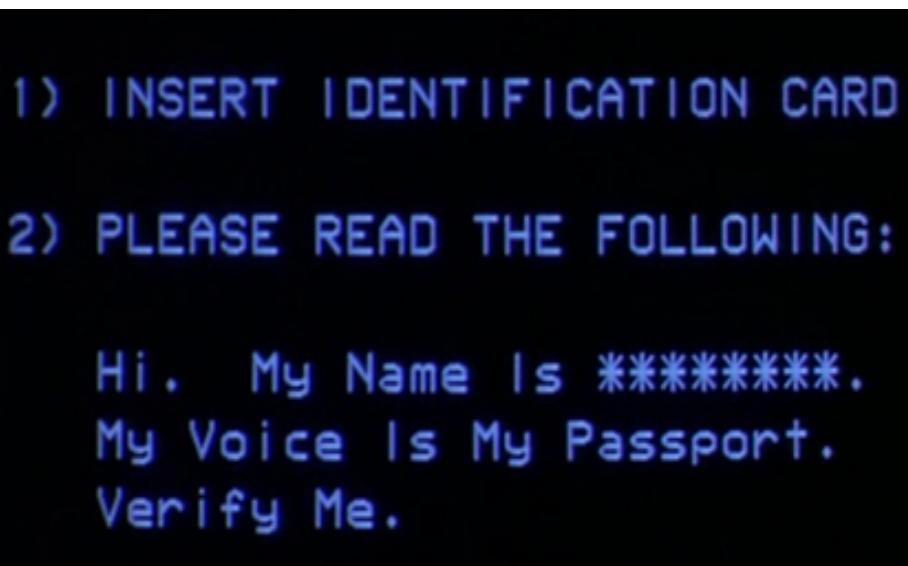
# Challenge: intra-speaker variation

Same speaker and same content but highly varied acoustics due to changes in style, voicing properties and other nuisances

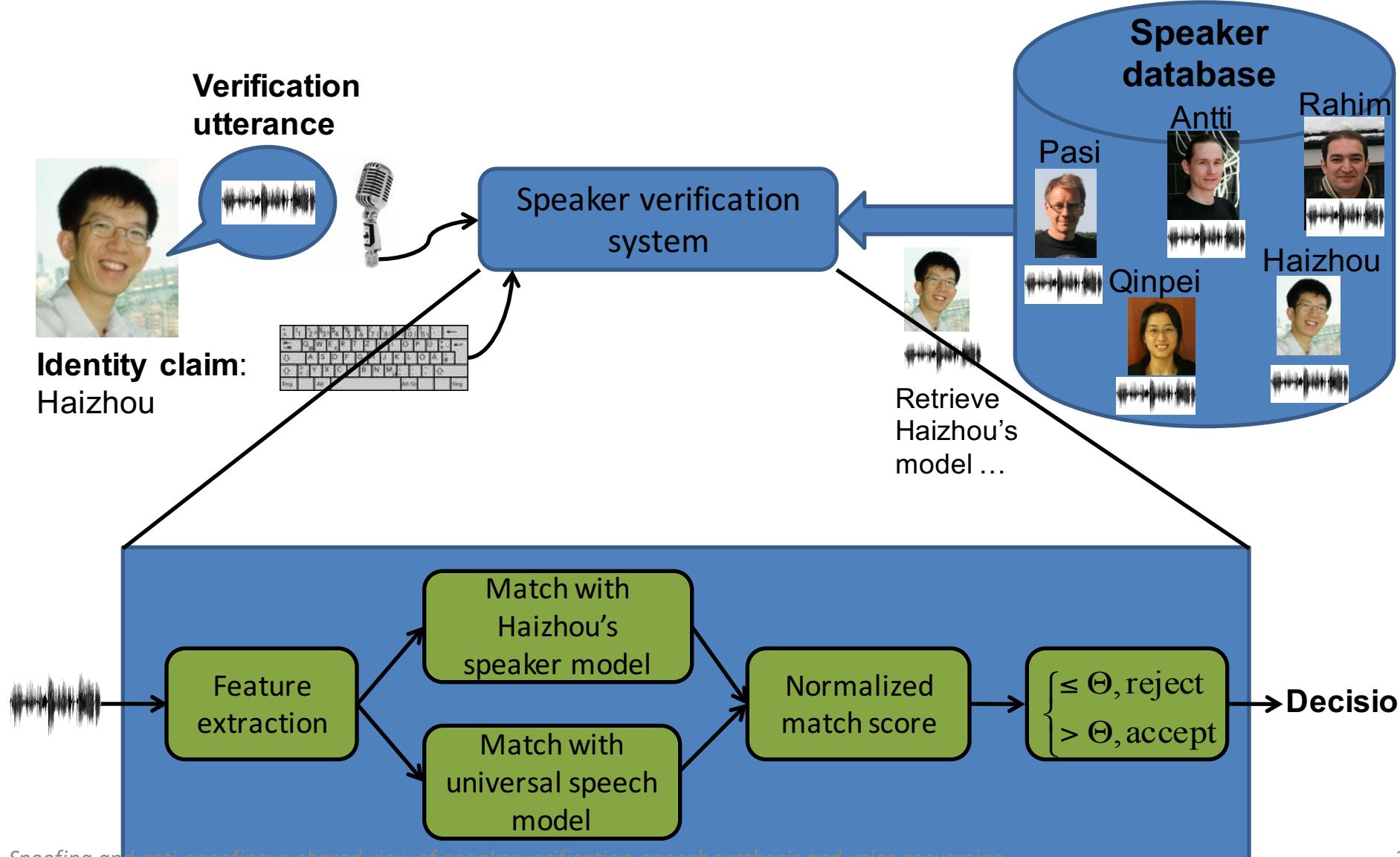


# Challenge: spoofing

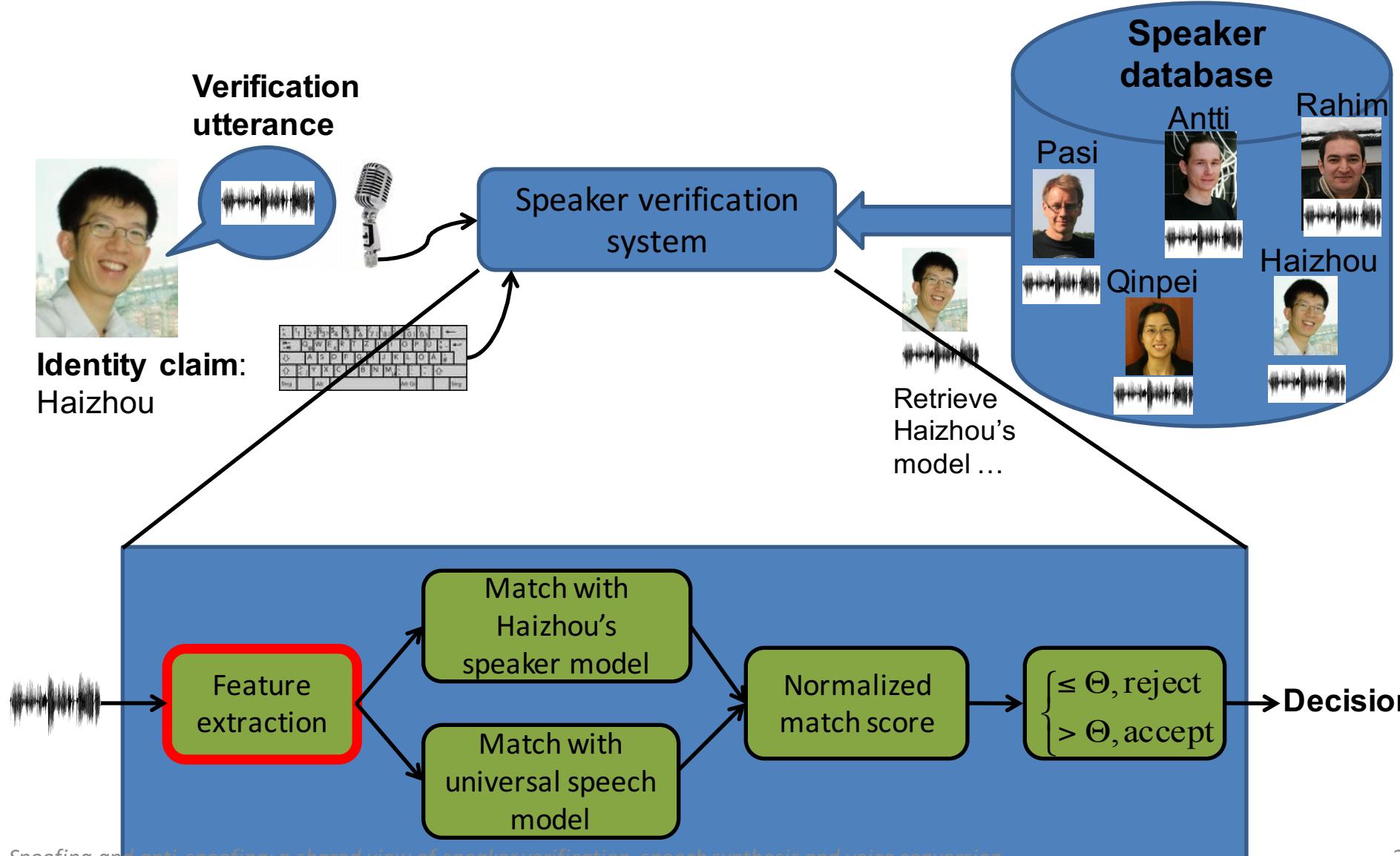
- Sneakers (1992)



# Internals of recognition system



# Feature extraction



## 《一剪梅》

红藕香残玉簟秋。  
轻解罗裳，独上兰舟。  
云中谁寄锦书来，  
雁字回时，月满西楼。  
花自飘零水自流。  
一种相思，两处闲愁。  
此情无计可消除，  
才下眉头，却上心头。

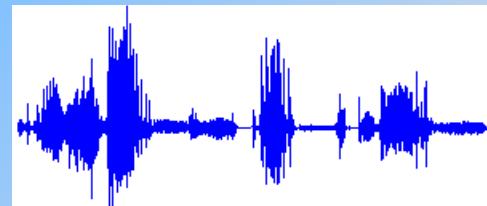


李清照

# Content

- Text-to-speech
- Speech-to-text
- ‘High-level’ speaker id

## Speech

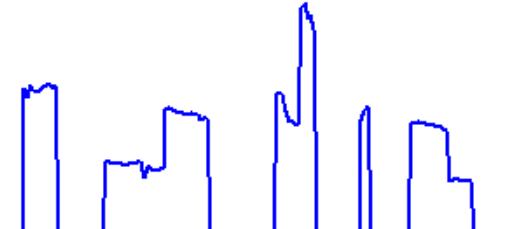


Shared domain of representation

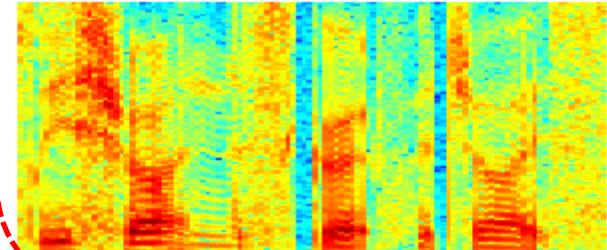
- Text-to-speech
- Voice conversion
- Automatic speaker verification

- Expressive TTS
- Singing synthesis
- Speaker verification

# Prosody



# Timbre



# Content

Phone, word or N-gram  
occupation counts

Categorical, sparse

- + Explicit/interpretable
- High error rate
- Difficult to extract

# Prosody

'Stylized' F0 / energy contours,  
Legendre polynomials,  
multinomial subspace models

# Timbre

Most practical (today)

Short-term spectrum: MFCC,  
LFCC, LPCC, PNCC ... with  $\Delta + \Delta^2$

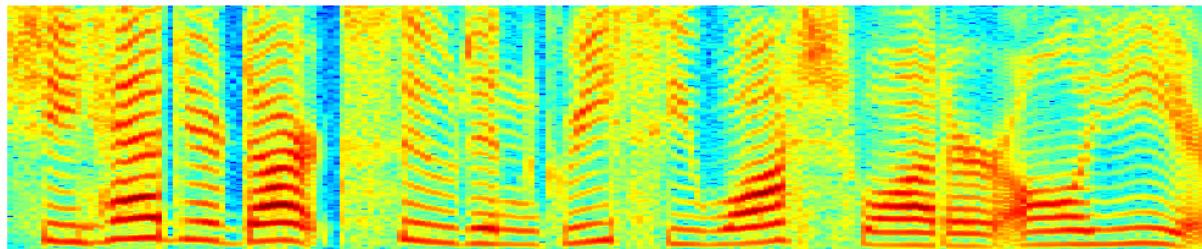
- + High accuracy
- + Easy & fast to compute
- Sensitive to disturbances

Continuous, dense

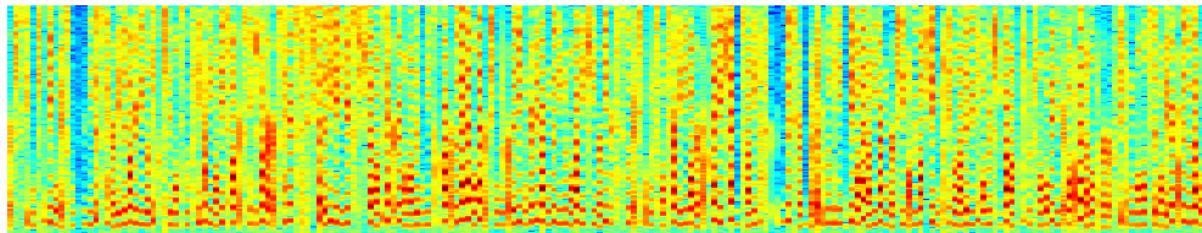
# Bag-of-frames illustrated

Ordering (temporal) information gets destroyed

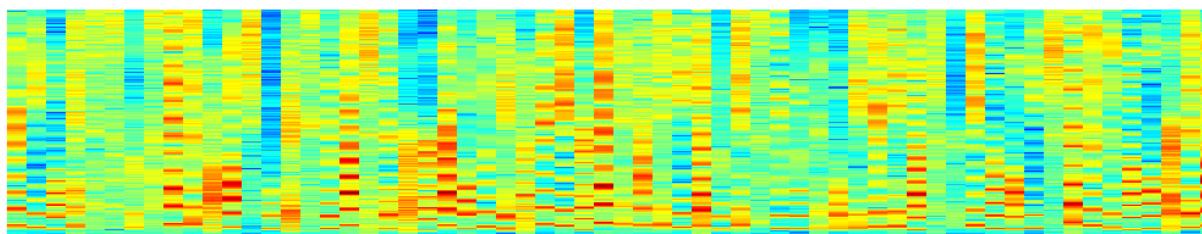
Original



Shuffled frames



Shuffled frames



Original



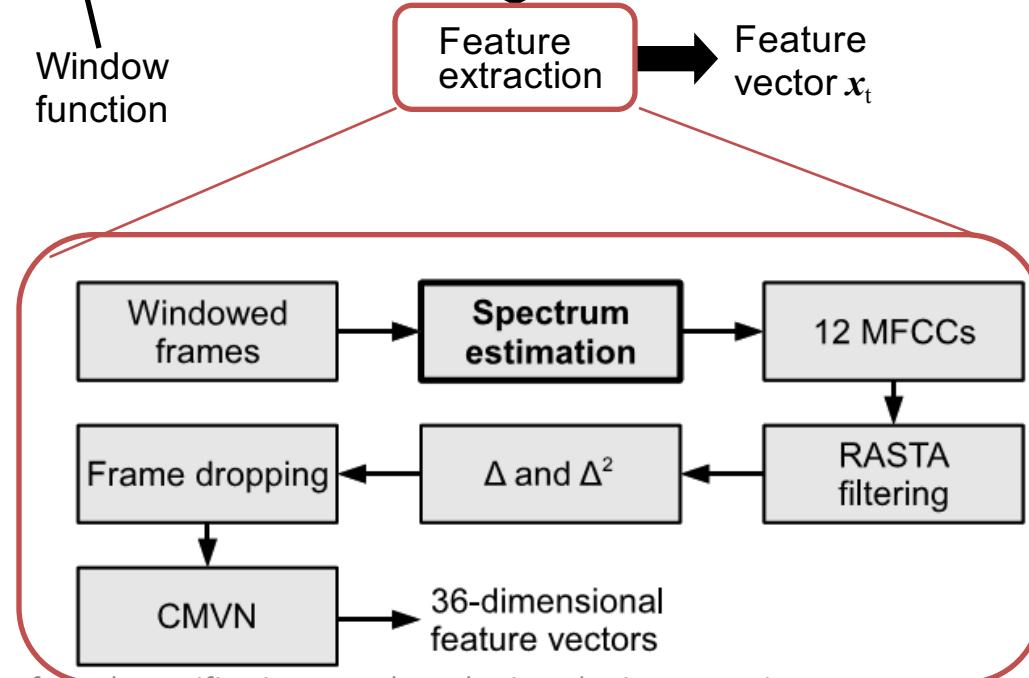
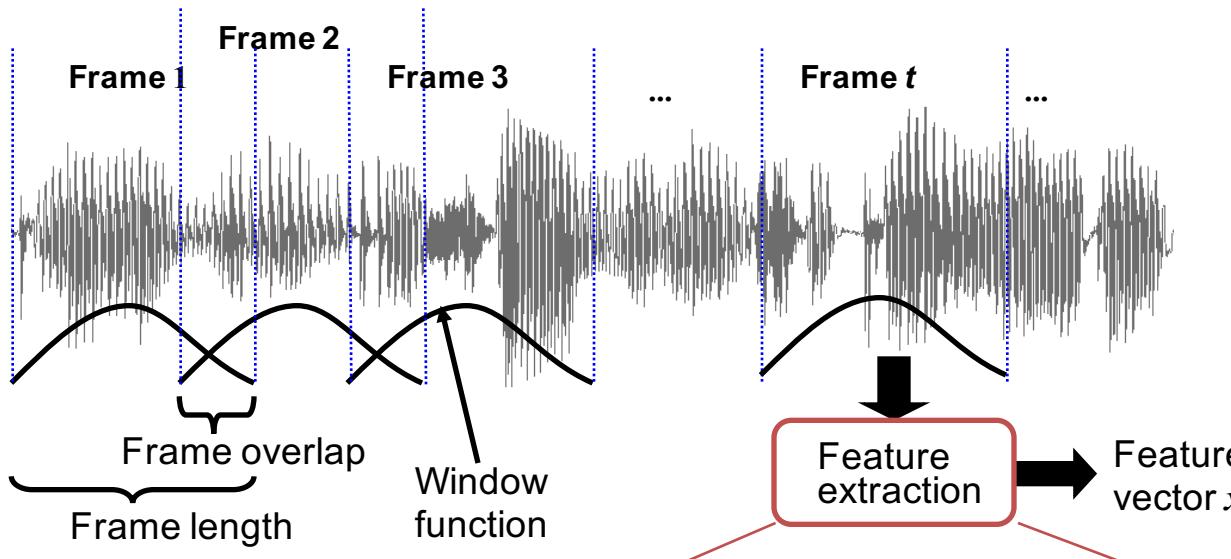
Frames shuffled,  
30 ms frame



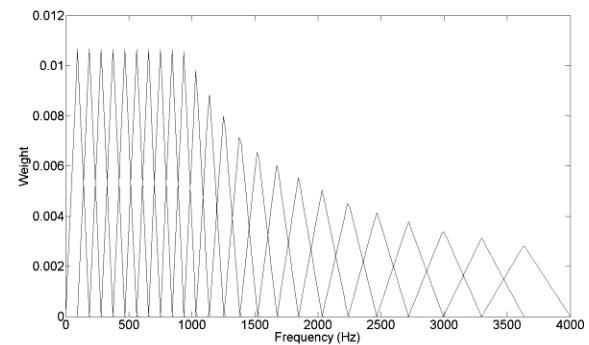
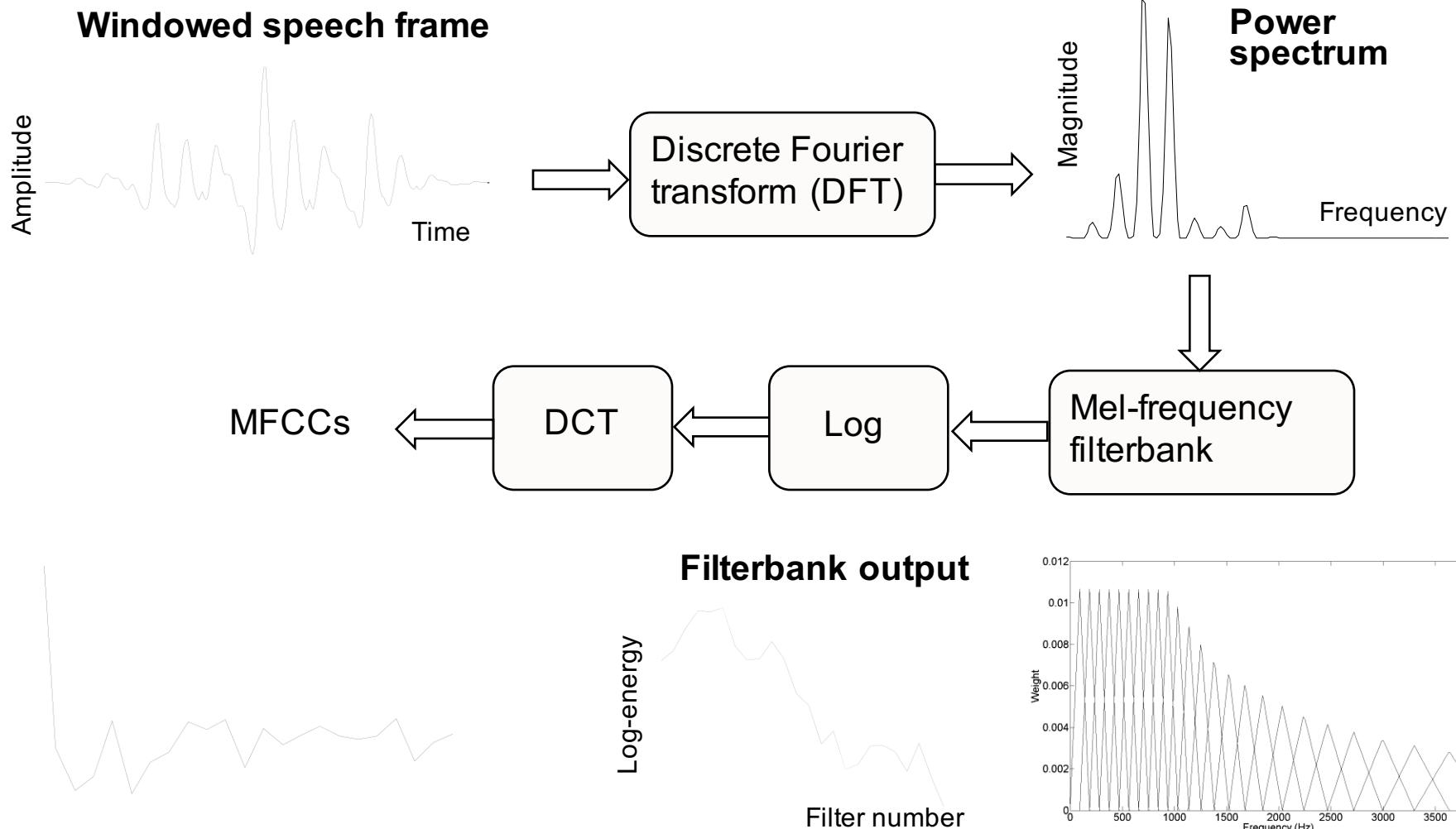
Frames shuffled,  
100 ms frame



# MFCC extraction

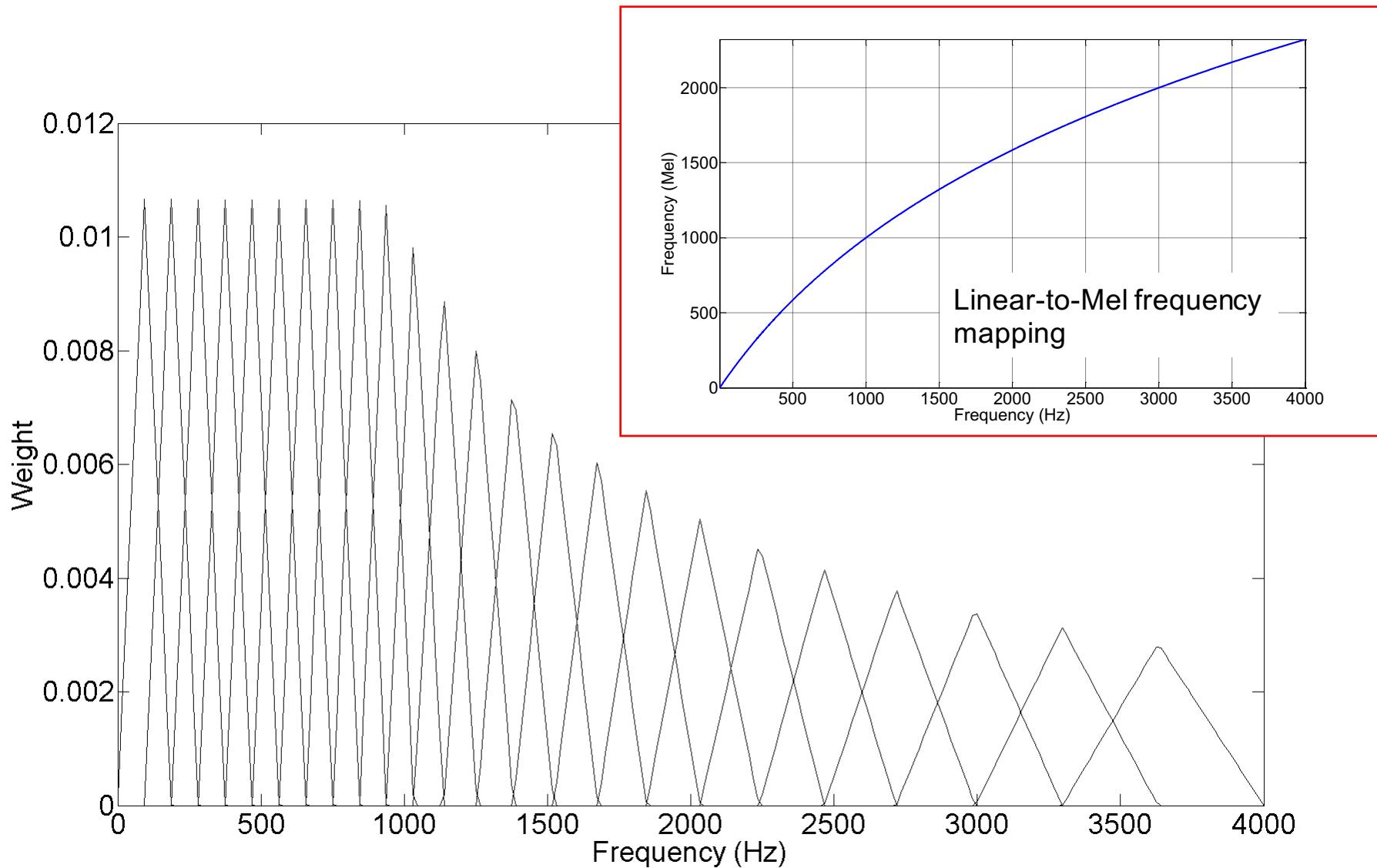


# Mel-frequency cepstral coefficient (MFCC) features



# Mel-frequency filterbank

[Generated using 'RASTAmat' package of Dan Ellis]

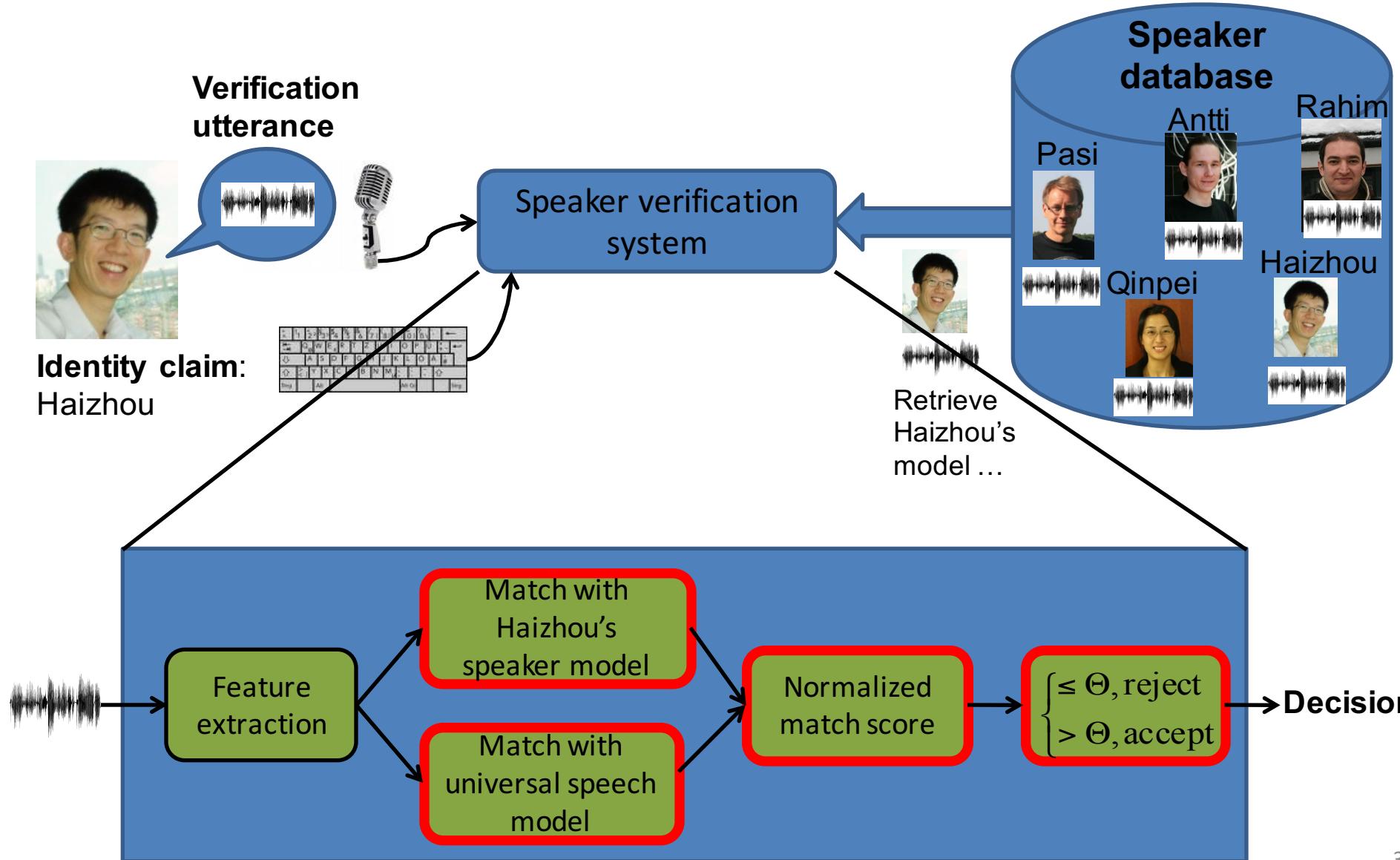


# Speaker verification: science and art of data-driven modeling

“... from the speaker-recognition research trend in the last decade, it seems that improving feature robustness beyond a certain level (for a variety of degradations) is extremely difficult—or, in other words, **data-driven modeling techniques have been more successful in improving robustness compared to new features**”

[John H.L. Hansen and Taufiq Hasan, Speaker Recognition by Machines and Humans: A Tutorial Review, IEEE Signal Processing Magazine, Nov 2015]

# Speaker modeling and comparison



# History of speaker modeling

## 1970s and before

- Long-term feature averaging

## 1980s and 1990s

- Dynamic time warping (DTW), vector quantiz. (VQ)
- Hidden Markov Models
- Early neural net models

- All rooted on GMMs
- 1996 onwards: NIST SREs
- Focus on text-independent models

## ~1995 to ~2005

- Gaussian mixture models (GMMs)
- Universal background model

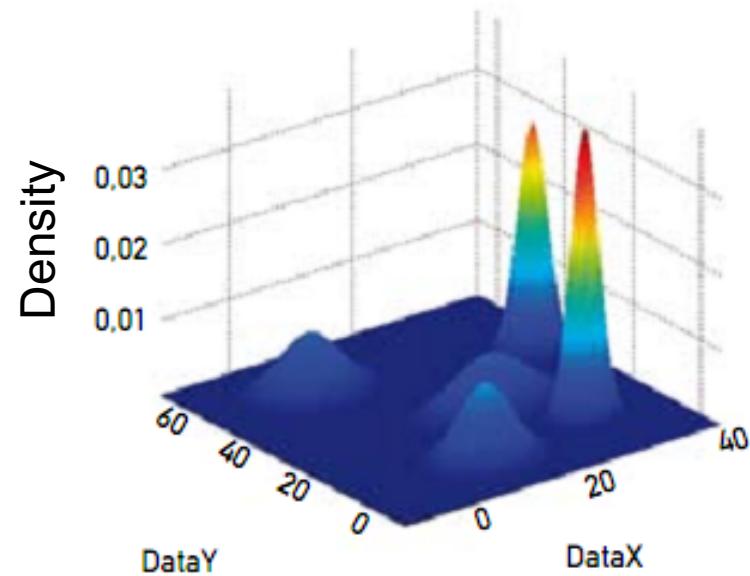
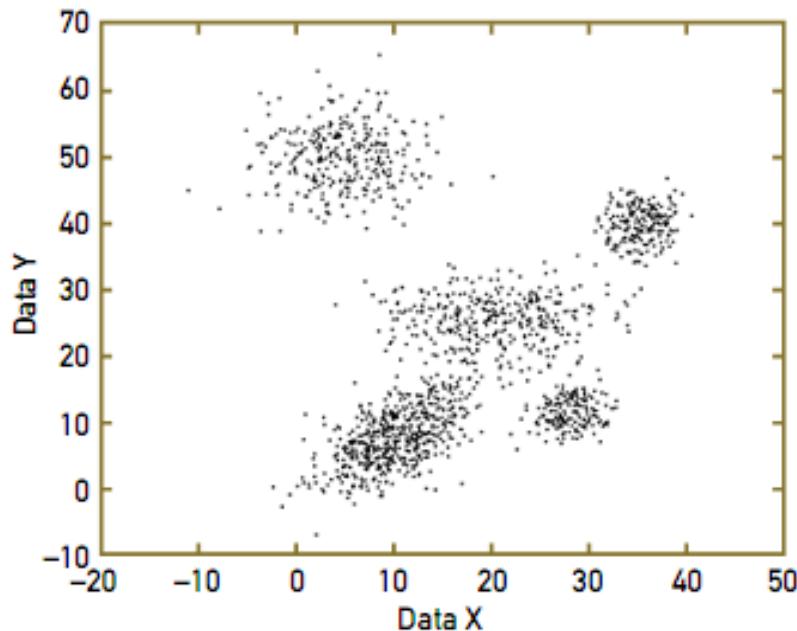
## 2005—today

- GMM supervectors
- Joint factor analysis (JFA)
- i-vectors
- Probabilistic linear discr. analysis (PLDA) scoring
- Deep neural nets

Pre-history

Modern era

# Gaussian mixture model (GMM)



$$p(\mathbf{x}|\lambda) = \sum_{k=1}^K P_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

↑      ↑      ↑  
Prior    Mean    Cov.  
probability   vector   matrix

$$\begin{aligned}P_k &\geq 0 \\ \sum_{k=1}^K P_k &= 1\end{aligned}$$

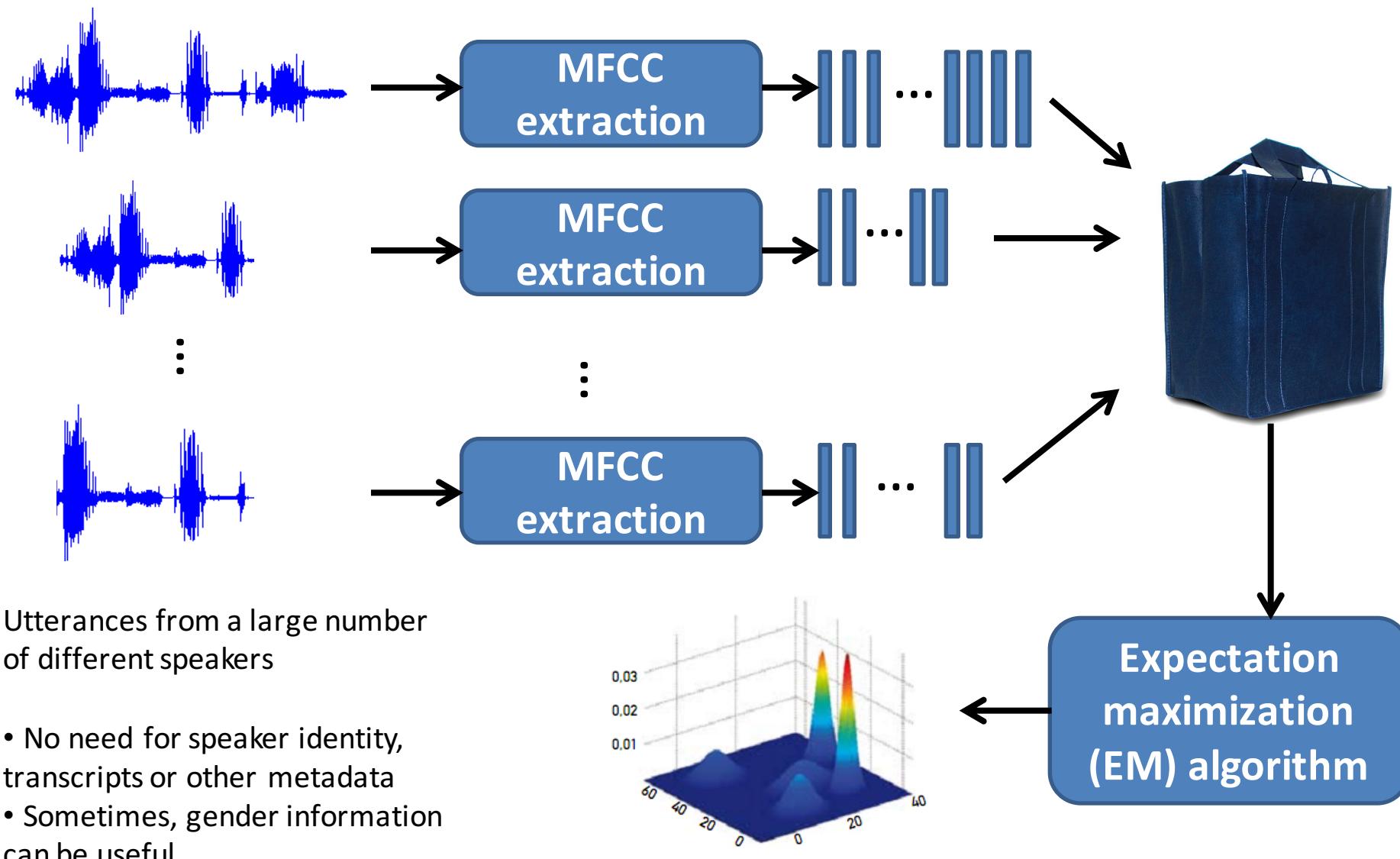
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\}$$

multivariate  
Gaussian density

# GMM-UBM

- $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$  is a sequence of test utterance feature vectors and  $\boldsymbol{\theta}_s$  is a GMM of speaker  $s$ , claimed to have 'generated'  $\mathbf{X}$
- A speaker verification system evaluates two hypotheses
  - $H_0$ : speaker  $s$  generated  $\mathbf{X}$
  - $H_1$ : anyone else but  $s$  generated  $\mathbf{X}$
- We evaluate
  - $p(\mathbf{X} | H_0) = p(\mathbf{X} | \boldsymbol{\theta}_s)$  target model likelihood
  - $p(\mathbf{X} | H_1) = p(\mathbf{X} | \boldsymbol{\theta}_{\text{ubm}})$  univ. background model likelihood
  - Log-likelihood ratio =  $\log p(\mathbf{X} | \boldsymbol{\theta}_s) - \log p(\mathbf{X} | \boldsymbol{\theta}_{\text{ubm}})$

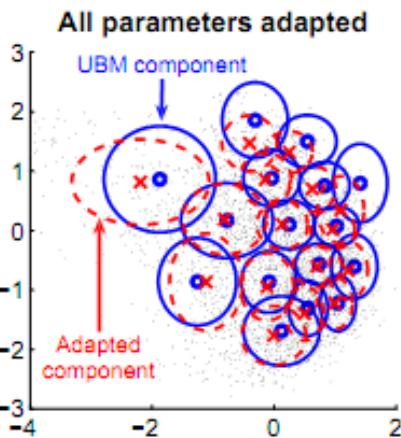
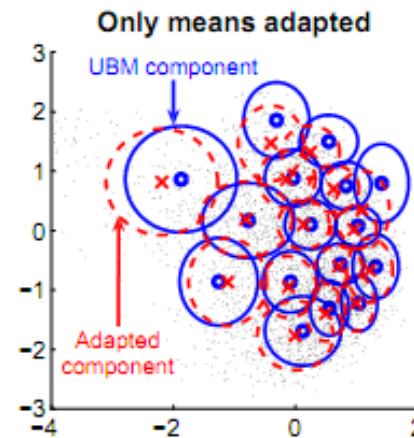
# Step 1: training the UBM



# Step 2: speaker enrollment via maximum a posteriori (MAP) adaptation

$$\boldsymbol{\mu}'_k = \alpha_k \tilde{\mathbf{x}}_k + (1 - \alpha_k) \boldsymbol{\mu}_k$$

↑  
 Adapted  
mean  
vector  
  
 ↑  
 Mean of  
training  
data  
  
 ↑  
 Prior mean  
from univ.  
background  
model (UBM)



$$\alpha_k = \frac{n_k}{n_k + r}$$

$$\tilde{\mathbf{x}}_k = \frac{1}{n_k} \sum_{t=1}^T P(k|\mathbf{x}_t) \mathbf{x}_t$$

$$n_k = \sum_{t=1}^T P(k|\mathbf{x}_t)$$

$$P(k|\mathbf{x}_t) = \frac{P_k \mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{m=1}^K P_m \mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}$$

Adaptation coefficient  
( $r$  = relevance factor,  
usually  $r = 16$ )

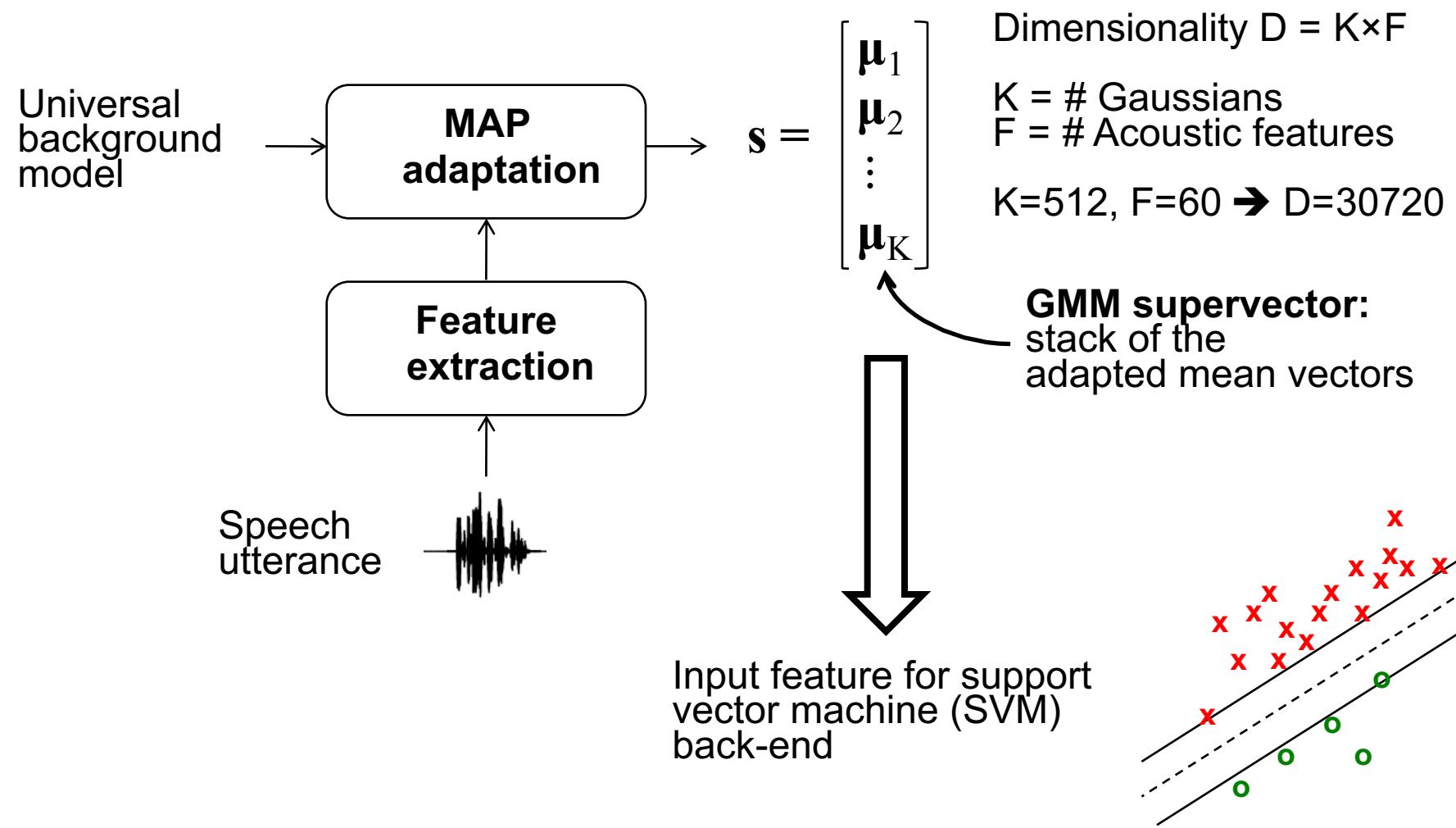
Mean of training data assigned  
to kth Gaussian

Soft count of vectors assigned  
to kth Gaussian

Posterior probability of the kth  
Gaussian for one feature  
vector

# GMM supervectors

[W. M. Campbell, D. E. Sturim, D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification", *IEEE Signal Proc Lett* 2006]



# Joint factor analysis (JFA) decomposition of the GMM supervector

$$\mathbf{S} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z}$$

The equation  $\mathbf{S} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z}$  is displayed with three blue curly braces underneath it. The first brace groups  $\mathbf{m}$  and  $\mathbf{V}\mathbf{y}$  and is labeled "Speaker". The second brace groups  $\mathbf{U}\mathbf{x}$  and is labeled "Channel". The third brace groups  $\mathbf{D}\mathbf{z}$  and is labeled "Residual".

- JFA model hyperparameters (trained in advance):  $\mathbf{m}$ : universal background model,  $\mathbf{V}$ : eigenvoice matrix,  $\mathbf{U}$ : eigenchannel matrix,  $\mathbf{D}$ : residual matrix
- Specific for an utterance:  $\mathbf{x}$  (channel factors),  $\mathbf{y}$  (speaker factors),  $\mathbf{z}$  (residual)
- “JFA cookbook” by Brno University of Techology (BUT)  
<http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo>

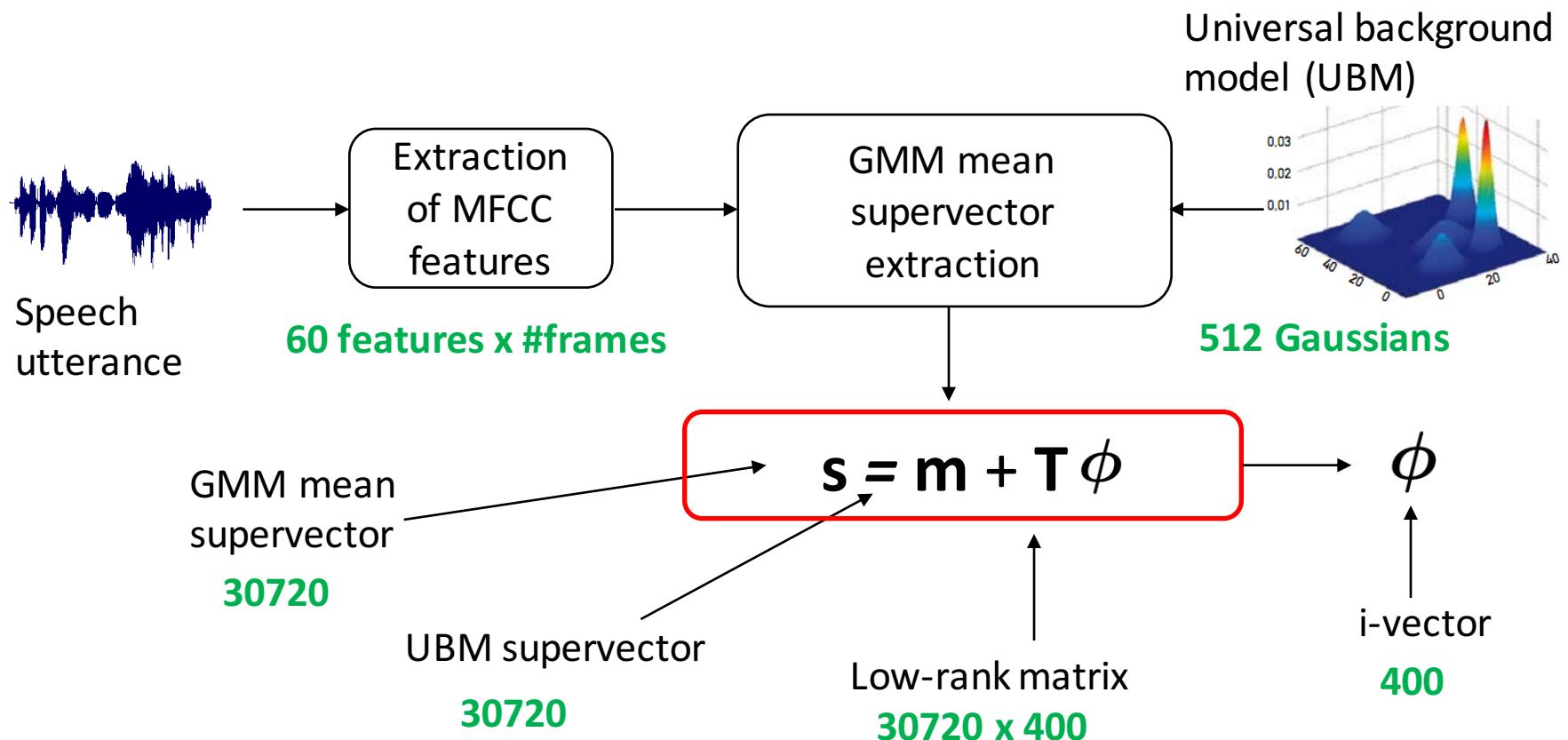
Kenny, P “Joint factor analysis of speaker and session variability : Theory and algorithms”, Technical report CRIM-06/08-13 Montreal, CRIM, 2005

Kenny, P., Boulian, G., Ouellet, P. and P. Dumouchel. “Joint factor analysis versus eigenchannels in speaker recognition”, *IEEE Transactions on Audio, Speech and Language Processing* 15(4), pp. 1435-1447, May 2007.

Kenny, P., Boulian, G., Ouellet, P. and P. Dumouchel. “Speaker and session variability in GMM-based speaker verification”, *IEEE Transactions on Audio, Speech and Language Processing* 15(4), pp. 1448-1460, May 2007.

# i-Vectors

Exactly the same training recipe as that of the eigenvoice matrix



[ N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011 ]

# i-vector normalization

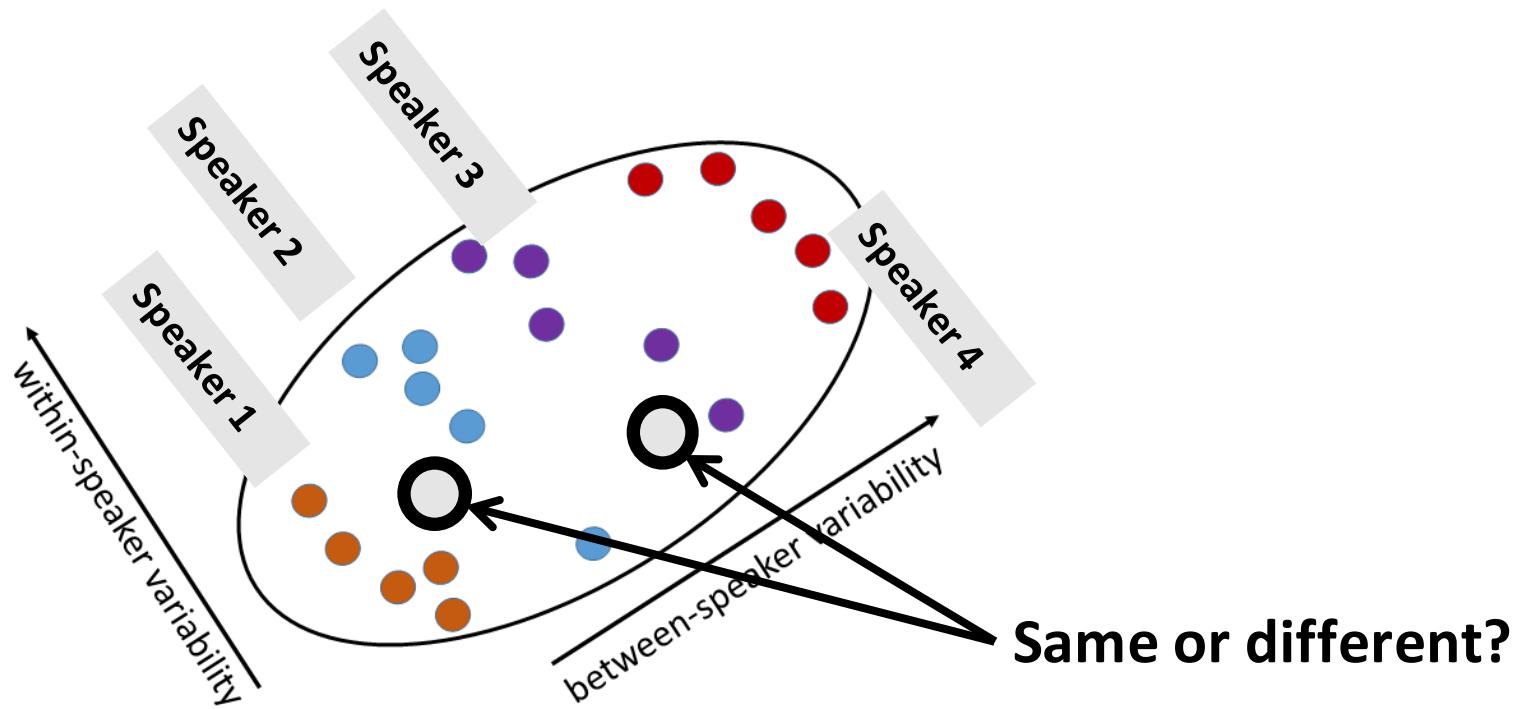
- I-vectors are ‘compressed’ GMM mean supervectors, based on MFCCs (or other spectral features), making ‘raw’ i-vectors sensitive to channel, noise and other variations
- **Within-class covariance normalization (WCCN)** [Hatch2006]
  - Introduced for normalization of GMM supervectors but useful for i-vectors as well
  - On a set of dev speakers, compute avg. within-speaker cov. matrix,  $\mathbf{C}$ , find the Cholesky decomposition of its inverse  $\mathbf{B}\mathbf{B}^T = \mathbf{C}^{-1}$ , apply  $\mathbf{B}^T\varphi_i$  on any i-vector
- **Linear discriminant analysis (LDA)**
  - Reduce dimensionality of i-vectors, using dev-speakers as classes
- **Length normalization** [Garcia-Romero 2011]
  - Project each i-vector to the unit sphere:  $\varphi_i \leftarrow \varphi_i / \|\varphi_i\|$
  - Useful for making i-vectors distributions closer to Gaussian

[Hatch2006] A.O. Hatch, S. Kajarekar, A. Stolcke, “Within-Class Covariance Normalization for SVM-based Speaker Recognition”, *Proc. Interspeech 2006*

[Garcia-Romero 2011] Daniel Garcia-Romero and Carol Y. Espy-Wilson, “Analysis of I-vector Length Normalization in Speaker Recognition Systems”, *Proc. Interspeech 2011*

# Probabilistic Linear Discriminant Analysis (PLDA)

[ S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE ICCV*, 2007, pp. 1–8 ]



$$\phi_{i,j} = \mu + \mathbf{V}\mathbf{y}_i + \mathbf{U}\mathbf{x}_{i,j} + \epsilon_{i,j}$$

$j$ :th i-vector of speaker  $i$

Bias

Between-speaker subspace  $\mathbf{V}$ , speaker factor  $\mathbf{y}_i$

Within-speaker subspace  $\mathbf{U}$ , factors  $\mathbf{x}_{i,j}$

Residual  $\sim N(\mathbf{0}, \Sigma)$

# Different flavors of PLDA

## Standard PLDA [Prince & Elder 2007]

$$\phi_{ij} = \mu + \mathbf{V}\mathbf{y}_i + \mathbf{U}\mathbf{x}_{ij} + \epsilon_{ij}$$

Two subspace models, **diagonal** cov. residual

## Simplified PLDA [Kenny 2010]

$$\phi_{ij} = \mu + \mathbf{S}\mathbf{y}_i + \epsilon_{ij}$$

One subspace model, **full** cov. residual

[Prince & Elder 2007] S.J.D. Prince and J.H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *IEEE 11th ICCV, pages 1–8, Oct 2007*

[Kenny 2010] P. Kenny. Bayesian speaker verification with heavy tailed priors. In *Proc. of the Odyssey Speak. and Lan. Recog. Workshop, Brno, Czech Republic, 2010.*

### Futher “two-covariance” PLDA variant:

[Brummer 2010] N. Brummer and E. De Villiers. The speaker partitioning problem. In *Proc. of the Odyssey Speak. and Lan. Recog. Workshop, Brno, Czech Republic, 2010.*

### Analysis and comparison of standard, simplified and two-cov. variants & scalable implementation:

[Sizov, Lee, Kinnunen 2014] Aleksandr Sizov, K-A Lee, T. Kinnunen, “Unifying Probabilistic Linear Discriminant Analysis Variants in Biometric Authentication”, *Proc. Joint Int. Workshop on Structural, Syntactic, and Statistical Pattern Recognition (S+SSPR 2014)*, pp. 464–475, Joensuu, Finland, August 2014

<https://sites.google.com/site/fastplda/>

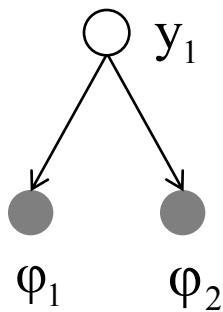
# PLDA scoring

Standard PLDA:  $\Phi_{ij} = \mu + \mathbf{V}\mathbf{y}_i + \mathbf{U}\mathbf{x}_{ij} + \boldsymbol{\varepsilon}_{ij}$

$$\text{score} = \log \frac{p(\varphi_1, \varphi_2 | H_0)}{p(\varphi_1, \varphi_2 | H_1)}$$

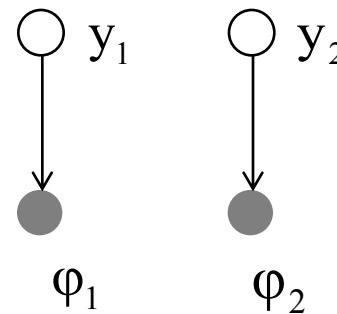
latent speaker factors

$H_0$  : same speaker



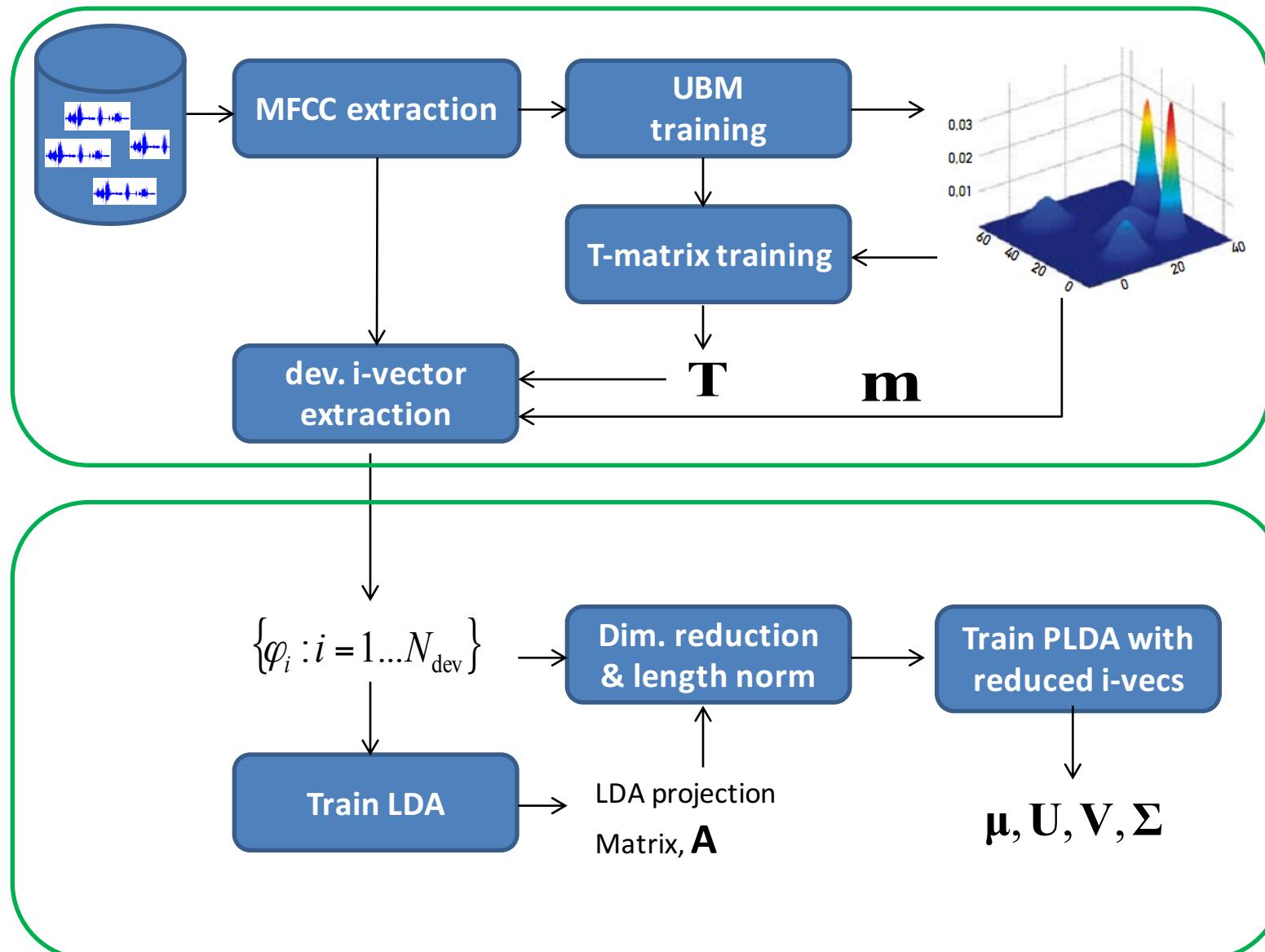
$$\begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix} = \begin{bmatrix} \mathbf{V} & \mathbf{U} & 0 \\ \mathbf{V} & 0 & \mathbf{U} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$$

$H_1$  : different speakers



$$\begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix} = \begin{bmatrix} \mathbf{V} & 0 & \mathbf{U} & 0 \\ 0 & \mathbf{V} & 0 & \mathbf{U} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$$

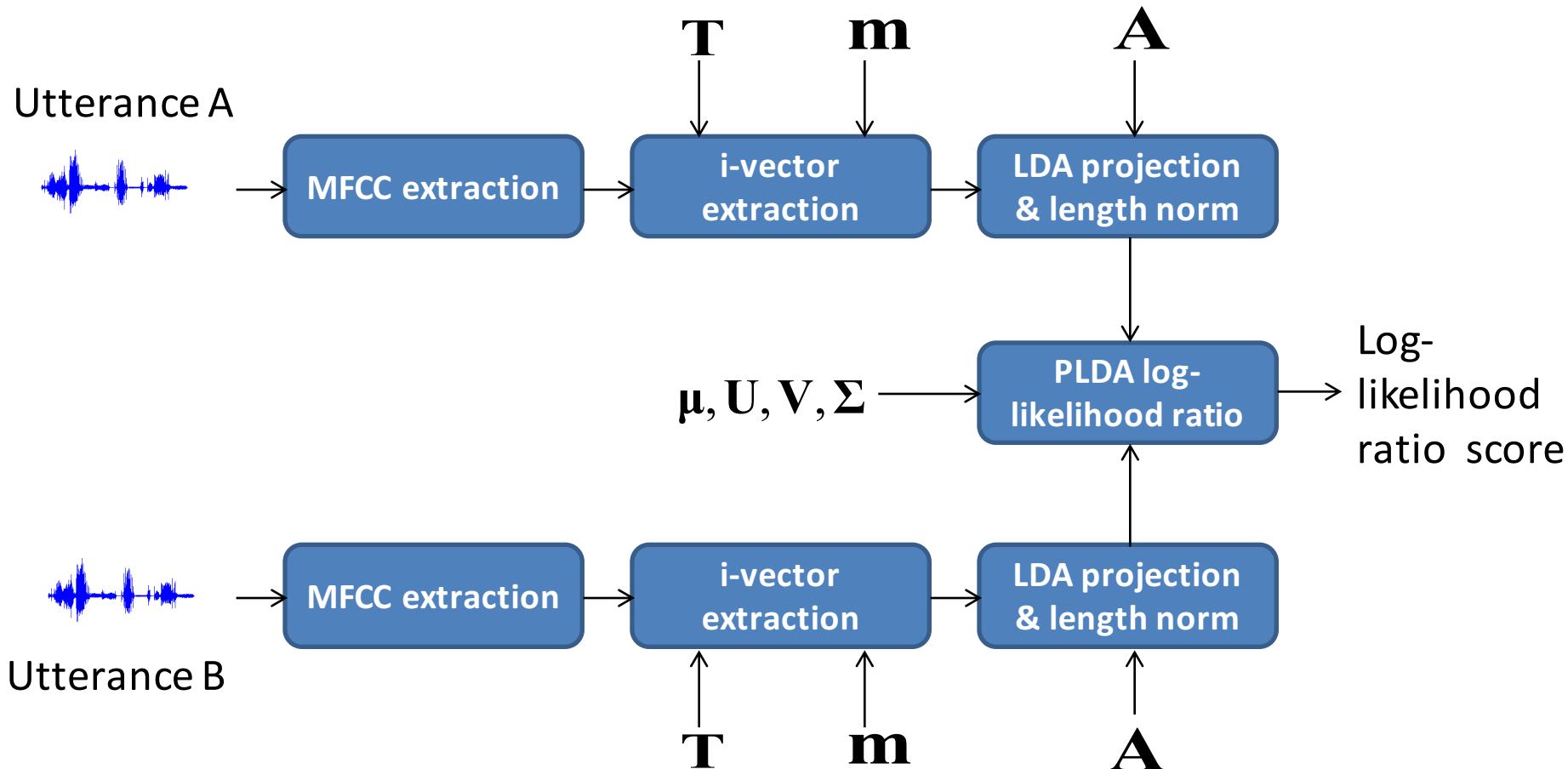
# i-vector PLDA: training



# i-vector PLDA: speaker verification

"Do A and B originate from same or different speakers ?"

The PLDA does not 'know' the speaker identity of neither one.



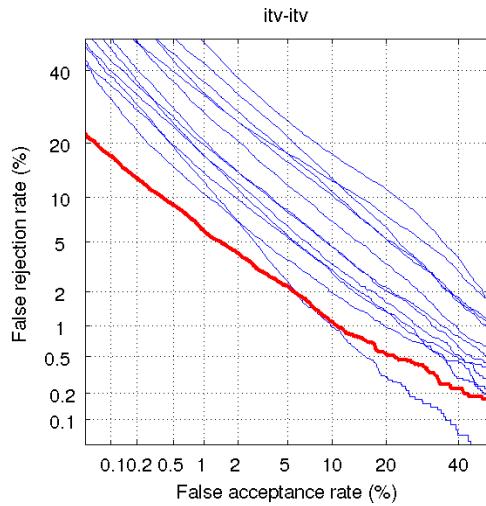
# Data to evaluate speaker recognition performance

- **Text-independent recognition:** benchmarks coordinated by National Institute of Standards and Technology (NIST) in the US, since 1996---
- **Text-dependent recognition:**
  - ‘RSR 2015’ corpus [Larcher et al 2014]
  - RedDots corpus [Lee & al 2015]

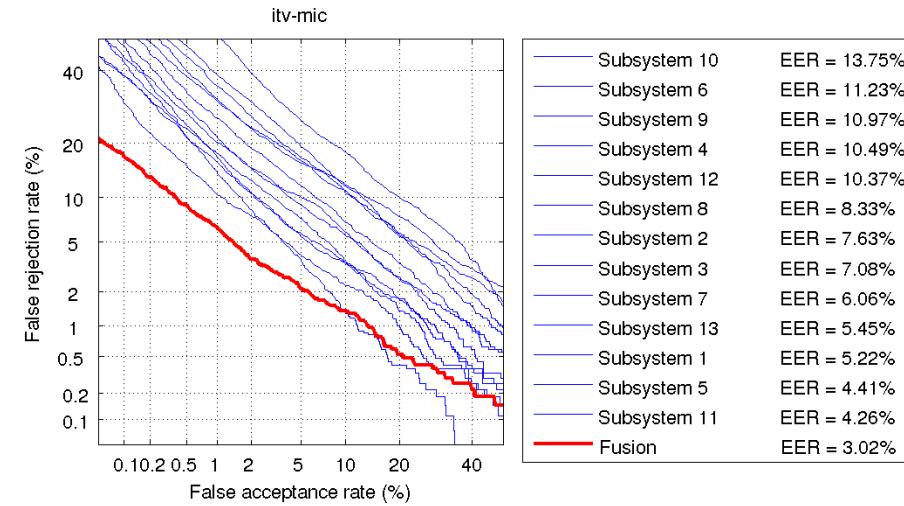
A. Larcher, KA Lee, B. Ma, H. Li, “Text-dependent speaker verification: Classifiers, databases and RSR2015”, Speech Communication 60, 56–77, 2014  
K.A. Lee et al, “The RedDots Data Collection for Speaker Recognition”, *Proc. Interspeech 2015*

# Performance measures:

## DET plots, DCF, EER

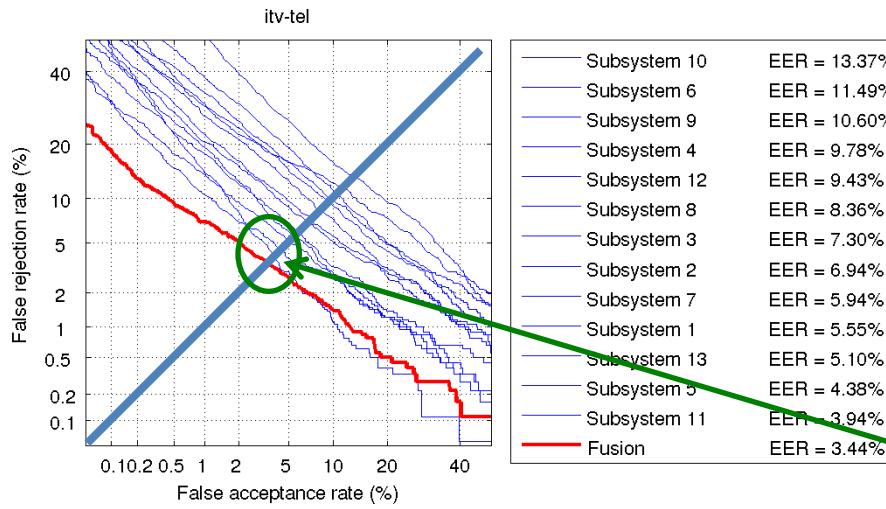


Subsystem 10	EER = 14.16%
Subsystem 6	EER = 11.51%
Subsystem 9	EER = 11.40%
Subsystem 12	EER = 10.35%
Subsystem 4	EER = 10.09%
Subsystem 8	EER = 8.16%
Subsystem 3	EER = 6.64%
Subsystem 2	EER = 6.56%
Subsystem 7	EER = 5.69%
Subsystem 1	EER = 5.26%
Subsystem 13	EER = 5.18%
Subsystem 5	EER = 4.12%
Subsystem 11	EER = 3.60%
Fusion	EER = 3.02%



$$DCF = C_{FR} P_{FR} P_{\text{target}} + C_{FA} P_{FA} (1 - P_{\text{target}})$$

$C_{FR} = 1$  Cost of false rejection  
 $C_{FA} = 1$  Cost of false acceptance  
 $P_{\text{target}} = 0.001$  Prior probability of the target speaker



**EER point:  $P_{fa} = P_{miss}$**

# Popular toolkits

Toolkit	Language
<b>ALIZE 3.0</b> <a href="http://www1.i2r.a-star.edu.sg/~alarcher/Softwares.html">http://www1.i2r.a-star.edu.sg/~alarcher/Softwares.html</a>	C++
<b>SPEAR Toolkit (based on BOB)</b> <a href="https://pypi.python.org/pypi/bob.spear/1.9.0">https://pypi.python.org/pypi/bob.spear/1.9.0</a> , <a href="http://idiap.github.io/bob/">http://idiap.github.io/bob/</a>	Python
<b>MSRidentity Toolbox</b> <a href="http://research.microsoft.com/en-us/downloads/2476c44a-1f63-4fe0-b805-8c2de395bb2c/">http://research.microsoft.com/en-us/downloads/2476c44a-1f63-4fe0-b805-8c2de395bb2c/</a>	Matlab
<b>Kaldi</b> <a href="http://kaldi.sourceforge.net/">http://kaldi.sourceforge.net/</a>	C++
<b>Sidekit</b> <a href="http://www-lium.univ-lemans.fr/sidekit/">http://www-lium.univ-lemans.fr/sidekit/</a>	Python

# ASV part: conclusions

- Most ASV systems use MFCCs or other short-term spectral features →
  - Sensitive to channel variation and noise
  - Vulnerable to spoofing with synthesis and voice conversion methods using similar features
- Extensive use of data-driven models
  - Classic systems:
  - universal background model (UBM)
  - Modern systems: UBMs, i-vector extractors, PLDA parameters...
- Highly active research community, thanks to common data sets!

# Outline

## Part 1

1. Introduction
2. Speaker verification
3. Speech synthesis
4. Voice conversion
5. Q&A

## Part 2

6. Spoofing
7. Countermeasures
8. ASVspoof 2015
9. Future
10. Q&A



# Various types of speech synthesisers

	Front-end	Back-end	Control of parameters
Formant synthesis (1970s)	Phonemes	Parametric model: Vocal tract model using formants	Handcrafted rules
Diphone synthesis (1980s)	Diphone	Concatenation of pre- recorded segments	Signal processing
Unit selection (1990s)	Context	Concatenation of pre- recorded segments	n/a
HMM synthesis (2000s)	Context	Parametric model: vocoder based on source-filter theory	Statistical model HMM

# 1990s

## Unit-selection synthesisers

# Unit selection synthesiser

- Conversion text to speech with larger database
  - sentence to **diphone with contexts**
  - search the optimal diphone unit sequence from database
  - Concatenate the selected diphone segments
- Unit selection synthesisers
  - 1990'
    - CHATR (Hunt and Black, ATR, Japan, '95)
    - Festival (Black, CSTR, Edinburgh, UK, '97)
    - AT&T Natural voice (USA)
- What is **context**?
- How is the search conducted?



*Examples: unit selection synthesizer*

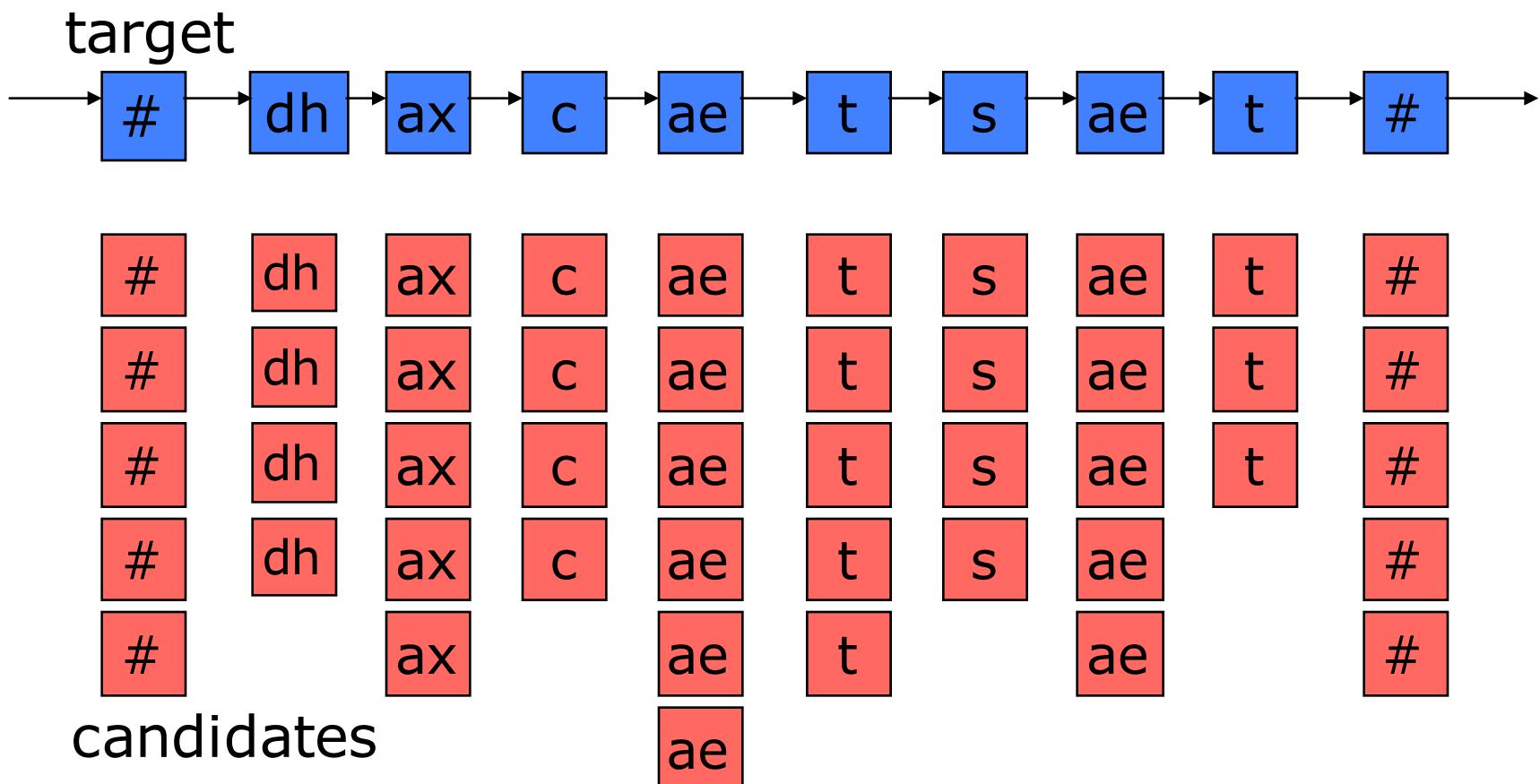
# Unit selection

- At synthesis time, if we can't find the speech sound from a precisely matching context, then choose a version of that sound from a **similar context**
  - in other words, a context that will have a **similar effect** on the sound
- For example:
  - can't find “phrase-final [a] in the context [n]\_[t]”
  - choose “phrase-medial [a] in the context [m]\_[d]”

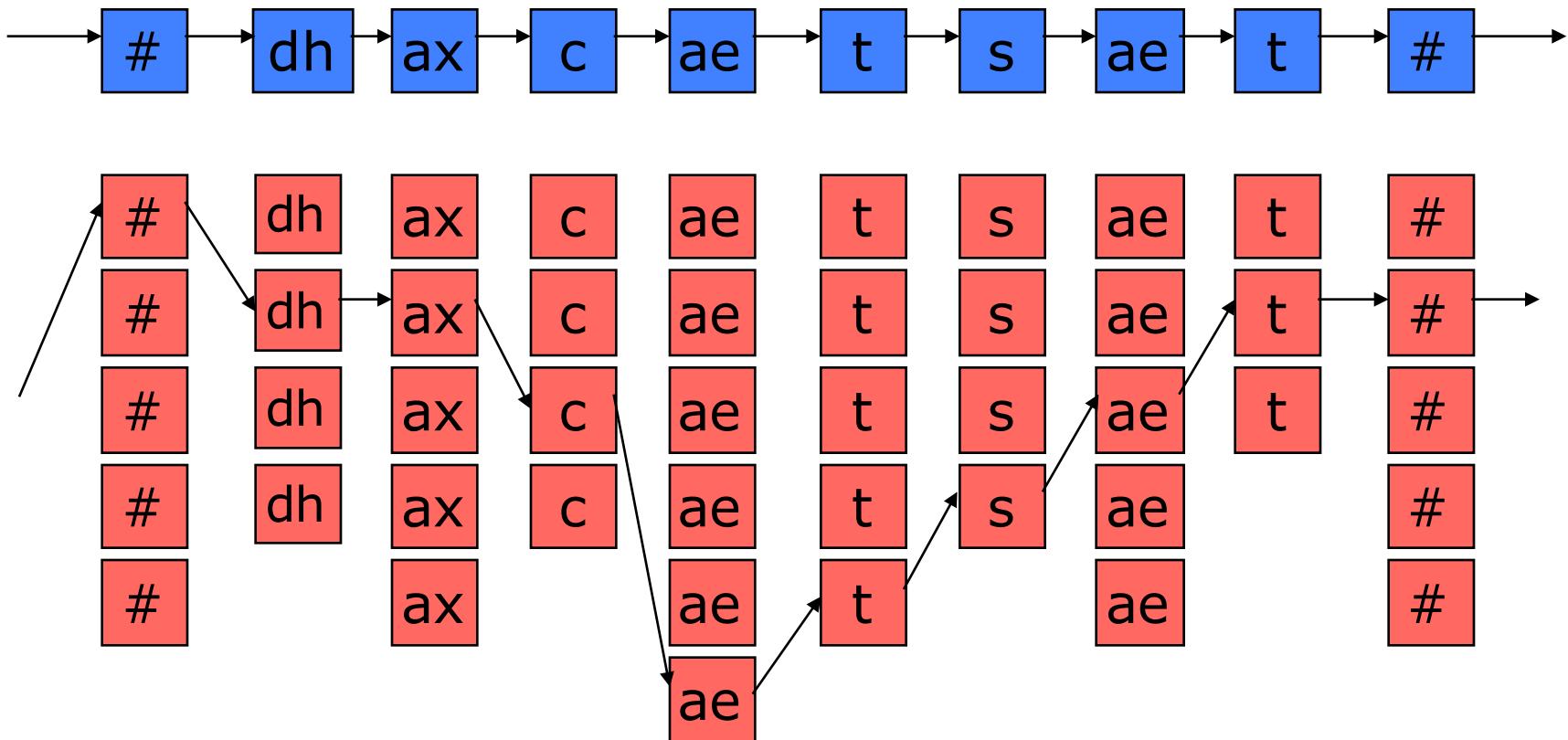
# Time aligned labels



# Many candidates (different contexts)



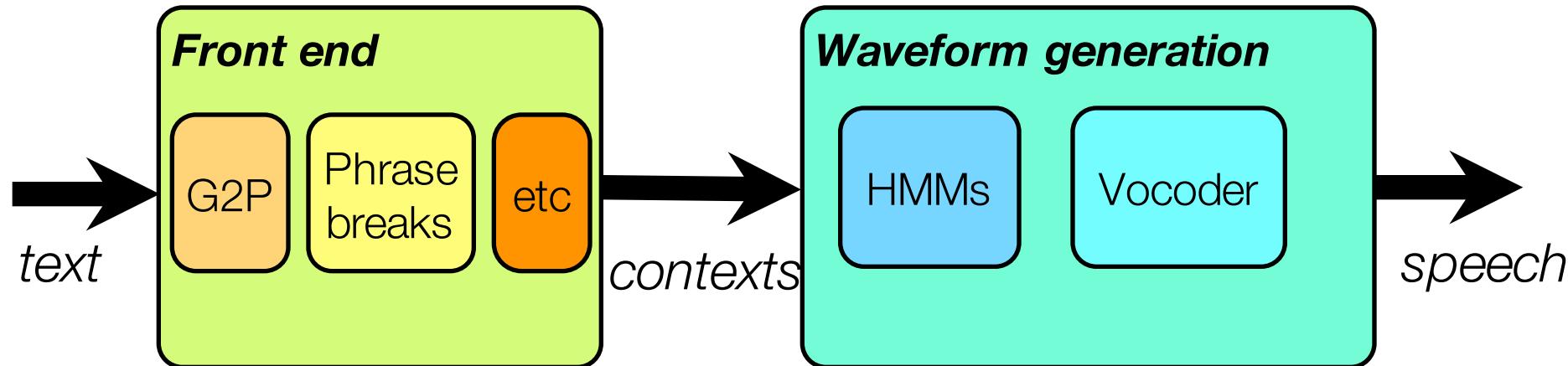
# Viterbi search



2000s

HMM-based speech synthesis

# HMM-based speech synthesis

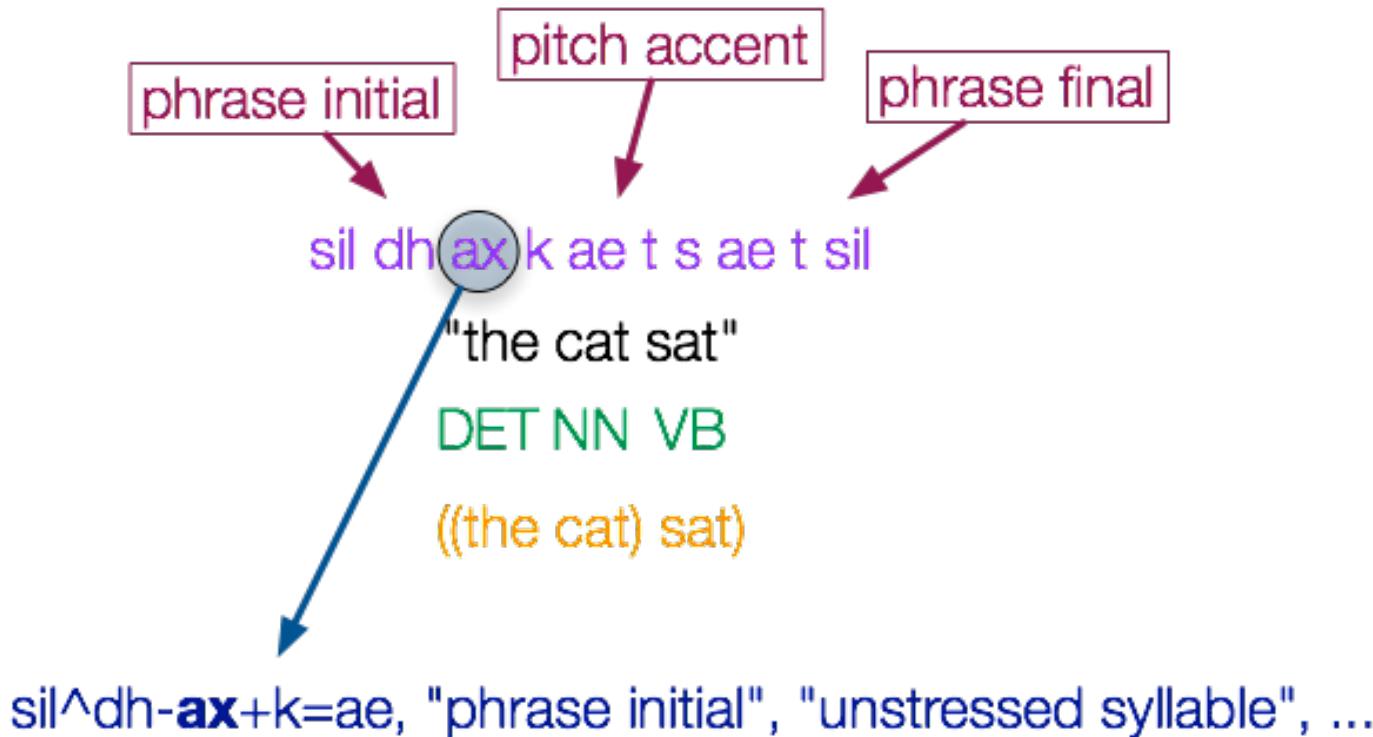


- Conversion text to speech with HMMs and vocoder
  - **Step 1:** Words to contexts
  - **Step 2:** Contexts determines HMMs to be used
  - **Step 3:** HMMs generate parameters required for vocoder
  - **Step 4:** Vocoder generates speech waveforms



# HMM-based speech synthesiser

## From words to contexts

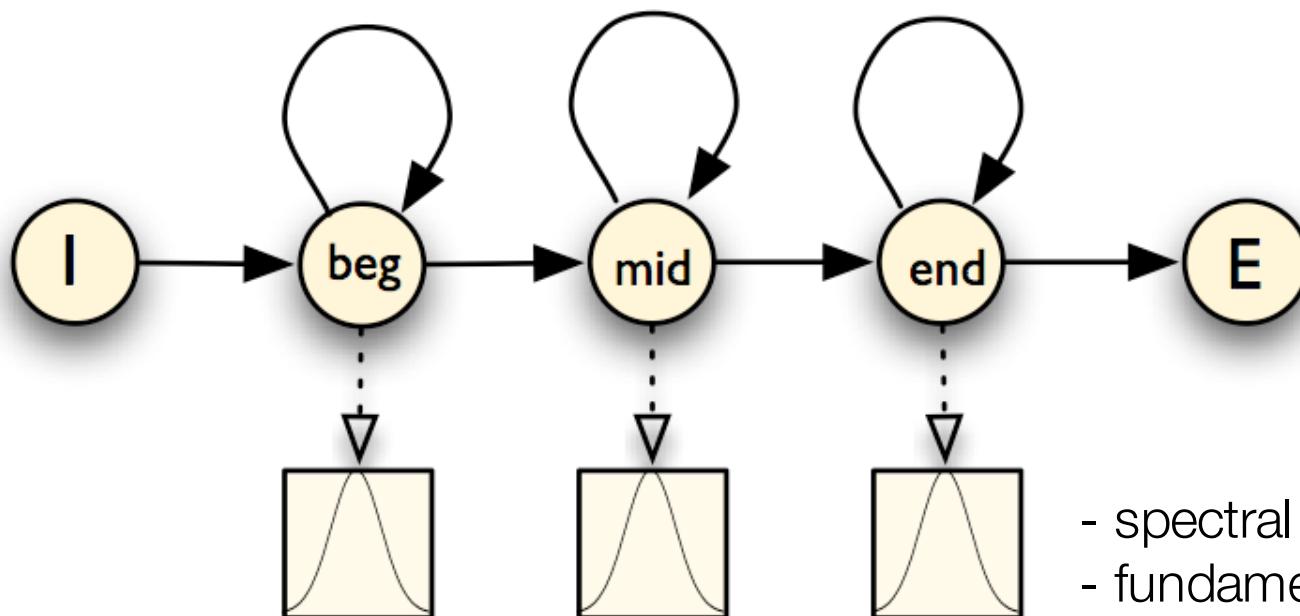


Tokuda, K.; Heiga Zen; Black, A.W. "An HMM-based speech synthesis system applied to English," Proceedings of 2002 IEEE Workshop on Speech Synthesis, pp. 227- 230, pp.11-13 Sept. 2002

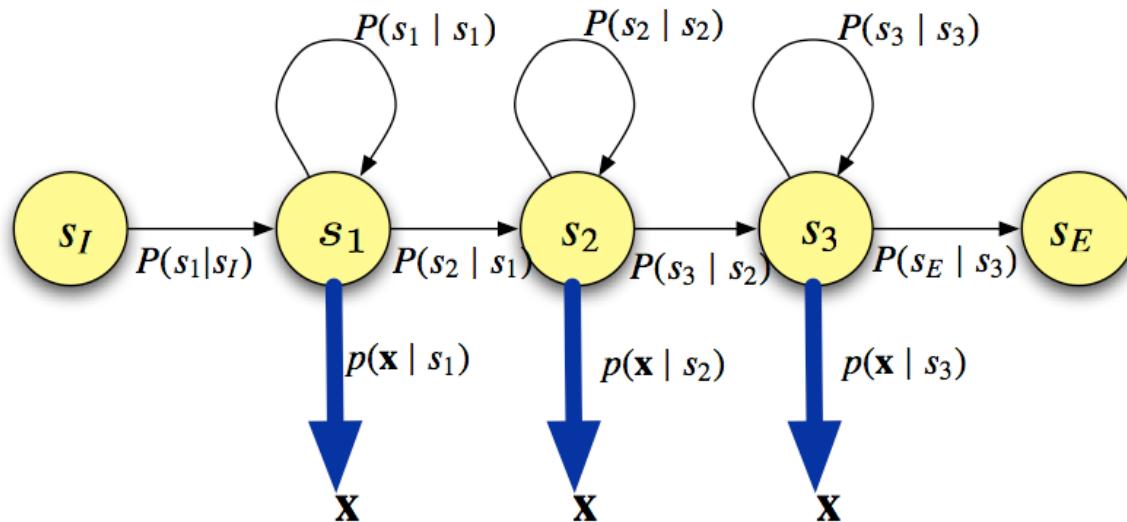
# HMM-based speech synthesiser

## From contexts to hidden Markov models

sil^dh-**ax**+k=ae, "phrase initial", "unstressed syllable", ...



# Output distribution



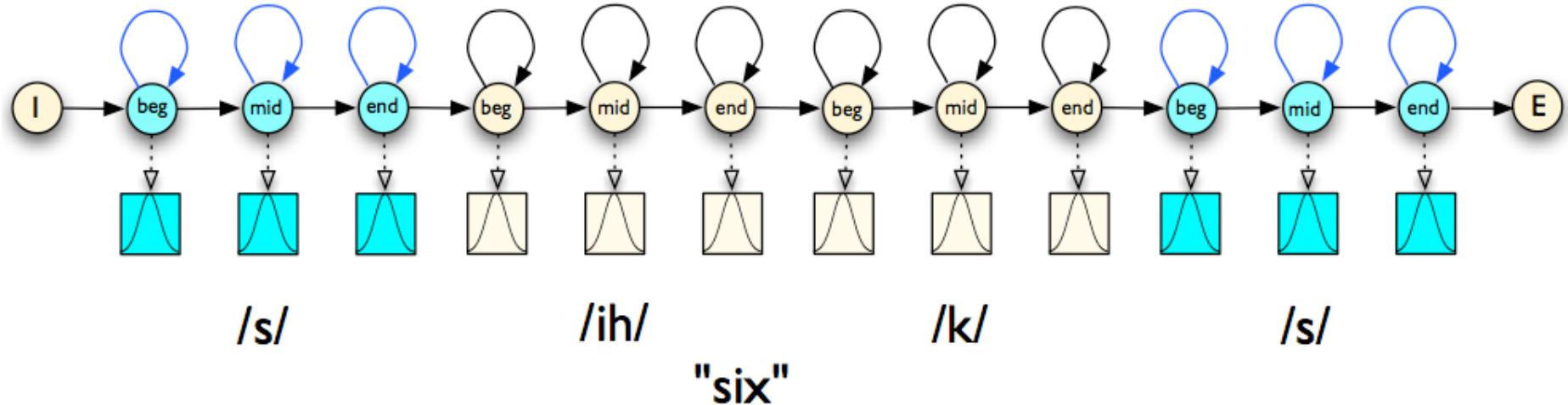
Single multivariate Gaussian with mean  $\mu^j$ , covariance matrix  $\Sigma^j$ :

$$b_j(\mathbf{x}) = p(\mathbf{x} | s_j) = \mathcal{N}(\mathbf{x}; \mu^j, \Sigma^j)$$

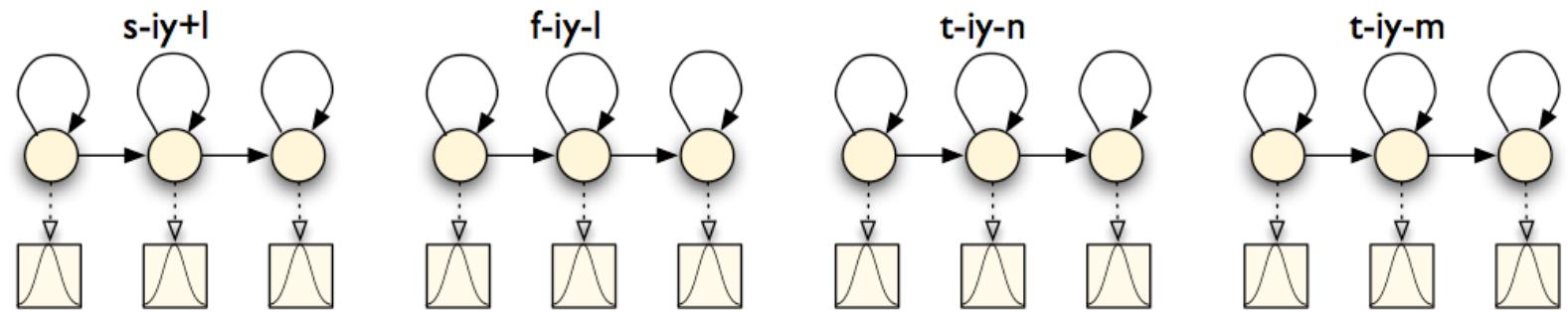
$M$ -component Gaussian mixture model:

$$b_j(\mathbf{x}) = p(\mathbf{x} | s_j) = \sum_{m=1}^M c_{jm} \mathcal{N}(\mathbf{x}; \mu^{jm}, \Sigma^{jm})$$

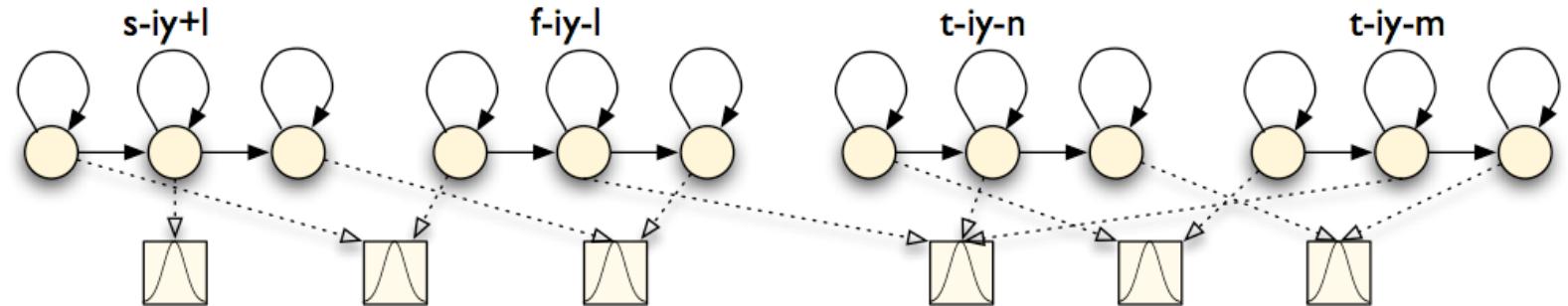
# Word model made from context-dependent phone HMMs



# State clustering/tying



Simple triphones (no sharing)

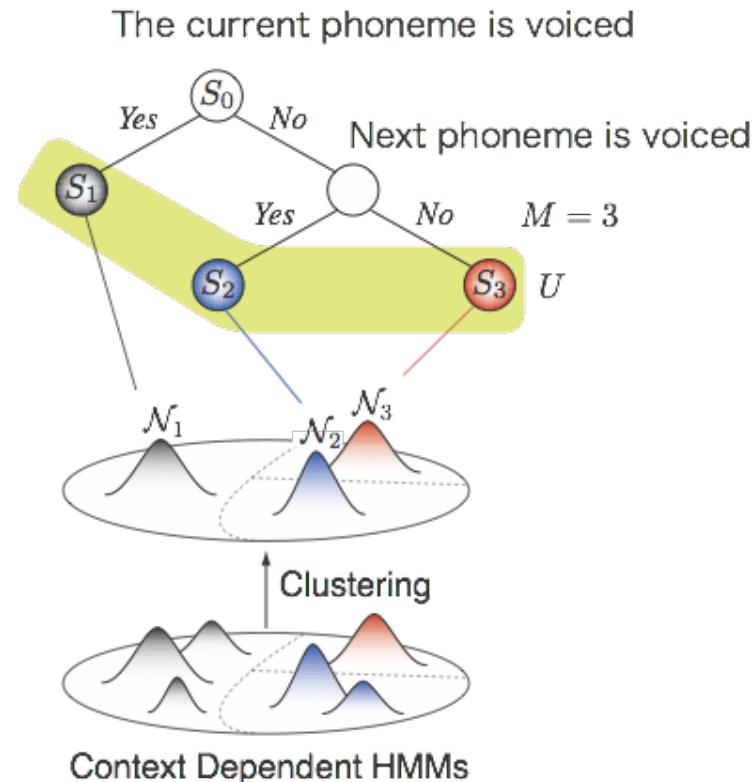


State-clustered triphones (state sharing)

# HMM-based speech synthesiser

## State tying of HMMs

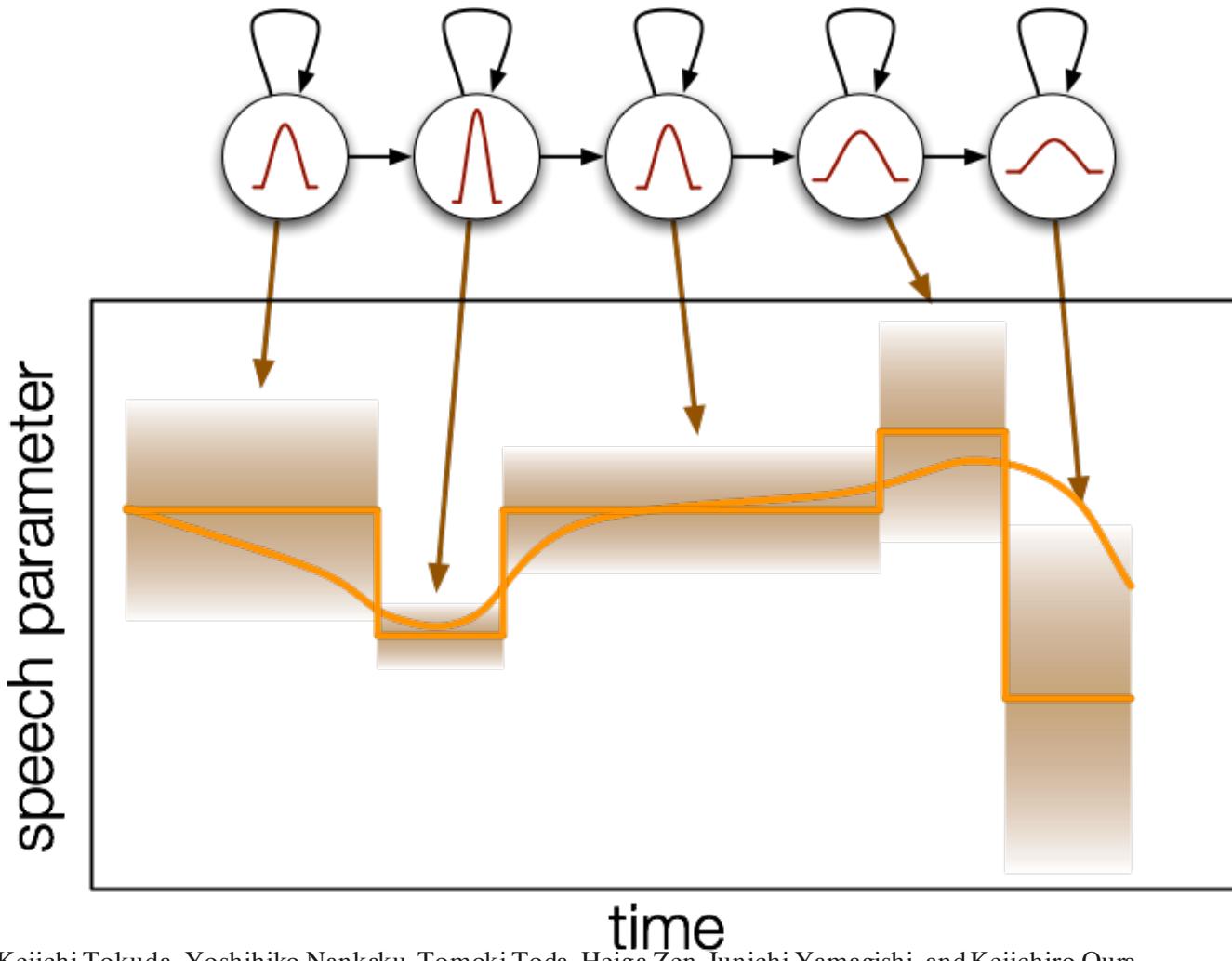
sil^dh-**ax**+k=ae, "phrase initial", "unstressed syllable", ...



S. J. Young, J. J. Odell, and P. C. Woodland. "Tree-based state tying for high accuracy acoustic modelling," Proceedings of the workshop on Human Language Technology (HLT '94) PA, USA, pp.307-312.1994

# HMM-based speech synthesiser

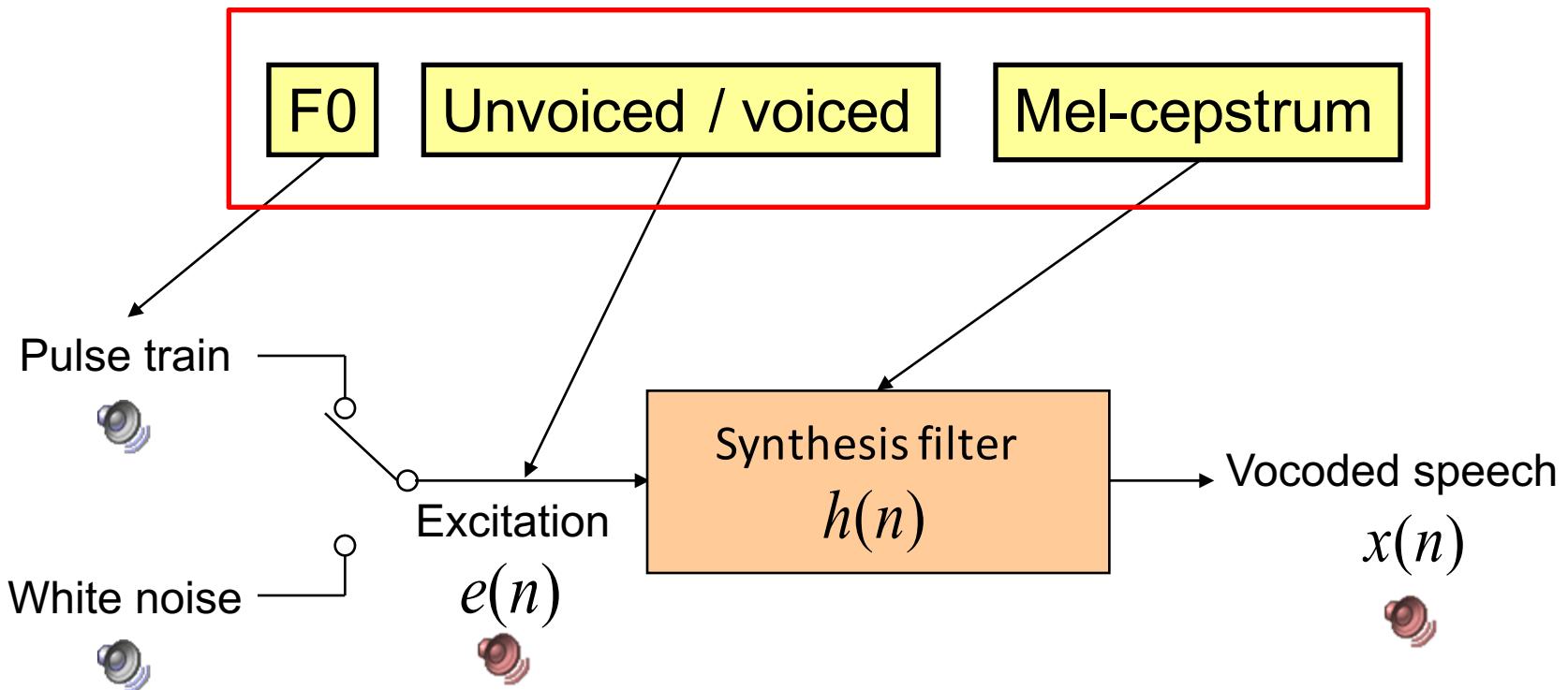
## Trajectory generation



Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura  
“Speech Synthesis Based on Hidden Markov Models” Proceedings of The IEEE, 2013

# Vocoder

Vocoder parameters generated from HMMs



Other vocoders (LSP, sinusoidal, Glottal, STRAIGHT, AHOCoder) can also be used

2005~

Adaptive HMM-based speech synthesis

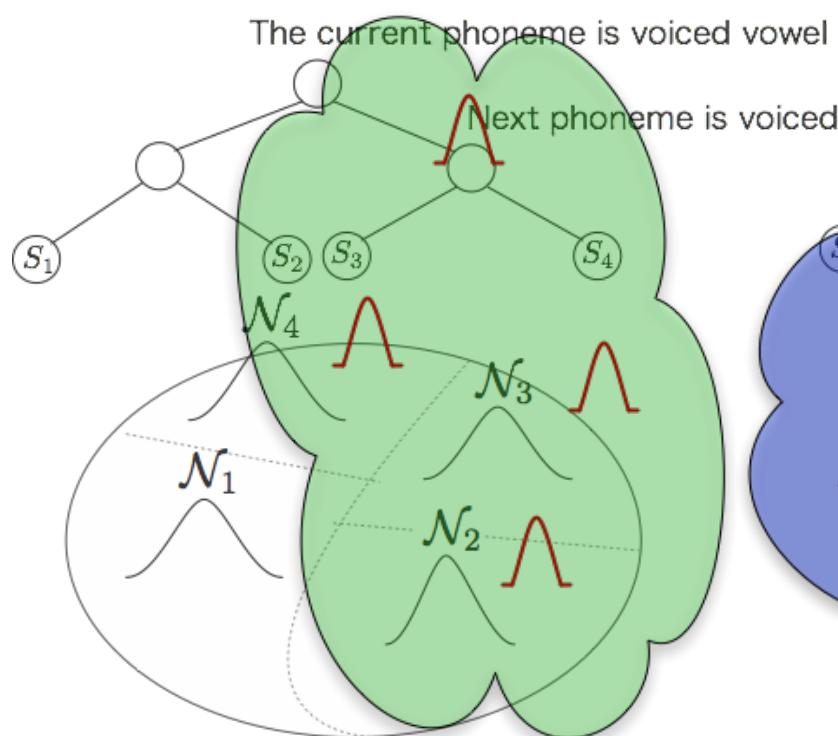
# Adaptation for speech synthesis

- One of the most important recent developments in speech recognition
- A **linear transform** is applied to every HMM parameter (Gaussian mean and variance) in order to adapt the model to new data
- Can be used to create new voices for speech synthesis:
  - Train HMMs on lots of data from multiple speakers
  - Transform the HMMs using a small amount of target speech
- This is a very exciting development in speech synthesis
- Provided data are available, any other acoustic difference can be adapted
  - speaker identity
  - emotion
  - dialect, and
  - the Lombard effect

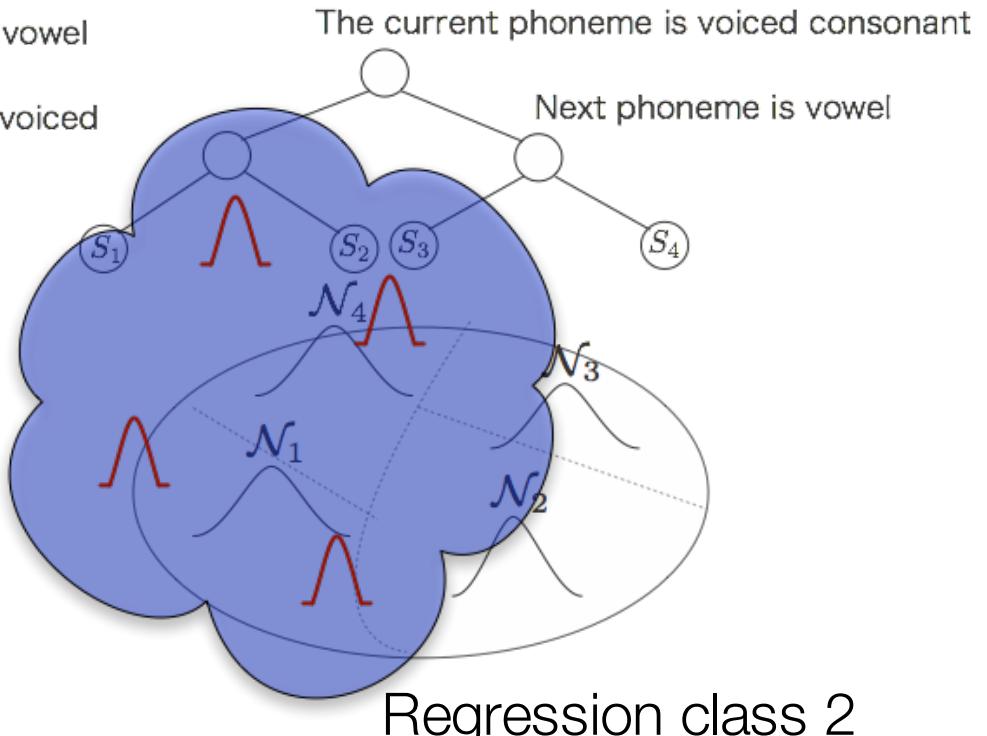
Yamagishi, J.; Kobayashi, T.; Nakano, Y.; Ogata, K.; Isogai, J.; , "Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm," IEEE Transactions on Audio, Speech, and Language Processing, , vol.17, no.1, pp.66-83, Jan. 2009

# Linear transforms of Gaussian pdfs of HMMs

Regression class 1



Consonants



Regression class 2

C.J. Leggetter, P.C. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech & Language, Volume 9, Issue 2, pp 171-185 April 1995,

# Adaptation to celebrity voices

Speech data can be acquired from broadcast, podcasts, lectures, telephone

Synthetic speech samples created in this scenario

George W Bush podcast:

Synthetic speech samples generated from HMMs adapted using speech data found on his podcasts

Sample



[Real-time demo \[web\]](#)

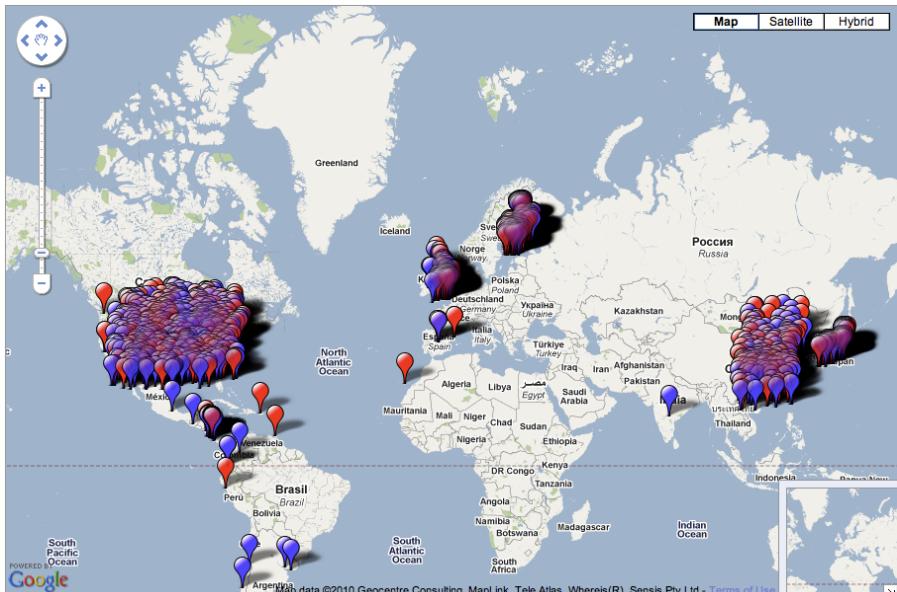
Queen Elizabeth-II's podcast

Synthetic speech samples generated from HMMs adapted using speech data found on her podcasts

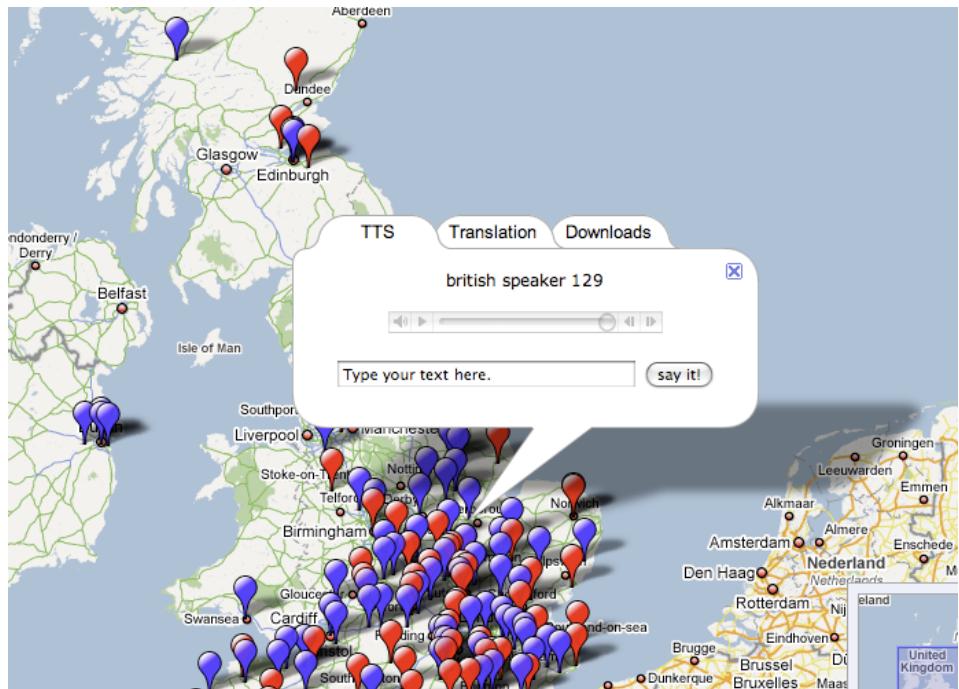
Sample



# Adaptation to individual voices in the world: Unlimited number of personalised TTS voices



<http://www.emime.org>



J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, M. Kurimo,  
“Thousands of Voices for HMM-based Speech Synthesis -- Analysis and Application of TTS Systems Built on Various ASR Corpora,”  
IEEE Trans. Audio, Speech, & Language Processing, vol.18, issue.5, pp.984-1004, July 2010

# Popular TTS toolkits

Toolkit	Language
<b>HTS Toolkit</b> <a href="http://hts.sp.nitech.ac.jp">http://hts.sp.nitech.ac.jp</a>	C
<b>HTS_engine_API</b> <a href="http://hts-engine.sourceforge.net">http://hts-engine.sourceforge.net</a>	C
<b>Flite</b> <a href="http://www.festvox.org/flite/">http://www.festvox.org/flite/</a>	C++
<b>Festival</b> <a href="http://www.cstr.ed.ac.uk/projects/festival/">http://www.cstr.ed.ac.uk/projects/festival/</a>	Scheme & C++
<b>OpenMARY</b> <a href="http://mary.dfki.de">http://mary.dfki.de</a>	Java

# Outline

## Part 1

1. Introduction
2. Speaker verification
3. Speech synthesis
- 4. Voice conversion**
5. Q&A

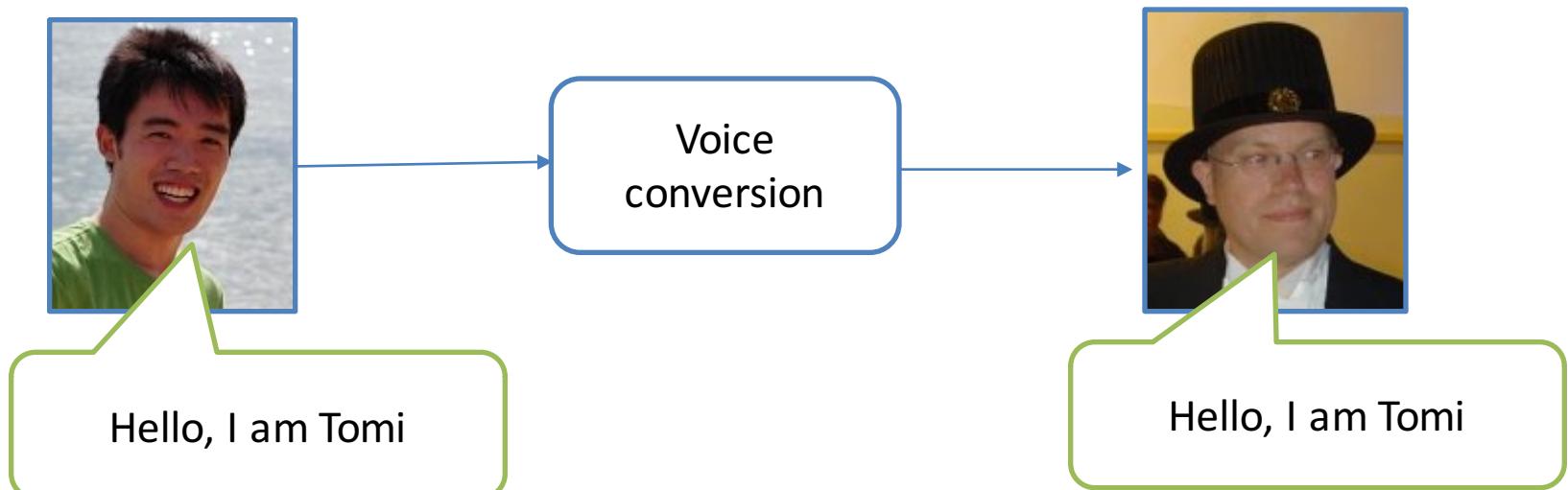
## Part 2

6. Spoofing
7. Countermeasures
8. ASVspoof 2015
9. Future
10. Q&A



# Voice conversion

- Converting para-linguistic information while keeping linguistic information unchanged
  - Para-linguistic information: speaker identity, speaking styles, etc



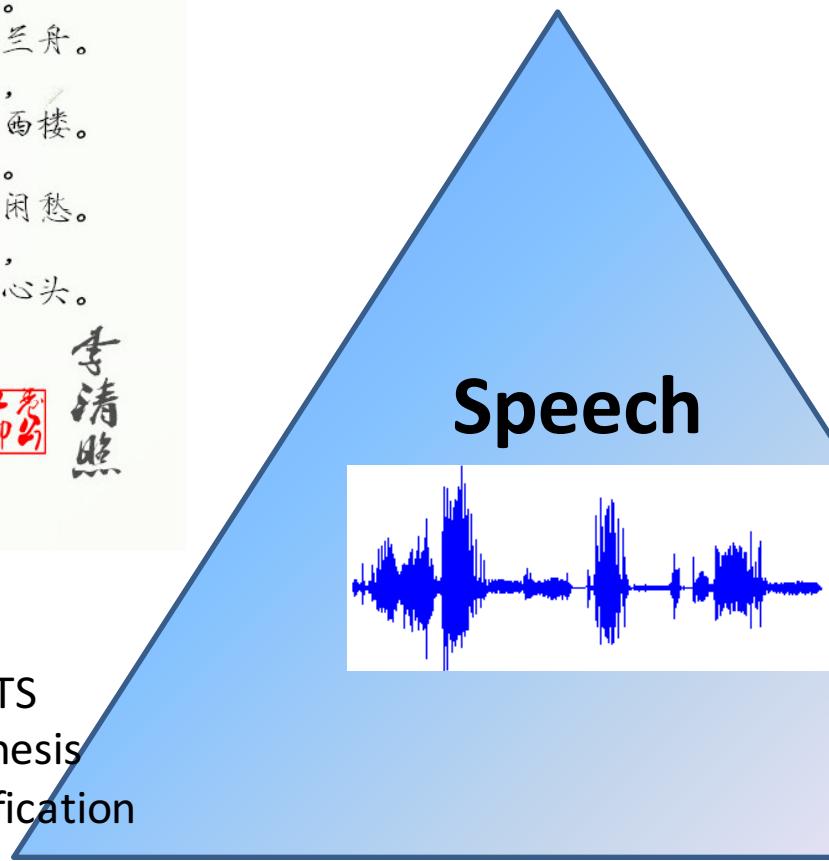
# 《一剪梅》

红藕香残玉簟秋。  
轻解罗裳，独上兰舟。  
云中谁寄锦书来，  
雁字回时，月满西楼。  
花自飘零水自流。  
一种相思，两处闲愁。  
此情无计可消除，  
才下眉头，却上心头。



# Content

- Text-to-speech
- Speech-to-text
- ‘High-level’ speaker id



- Expressive TTS
- Singing synthesis
- Speaker verification

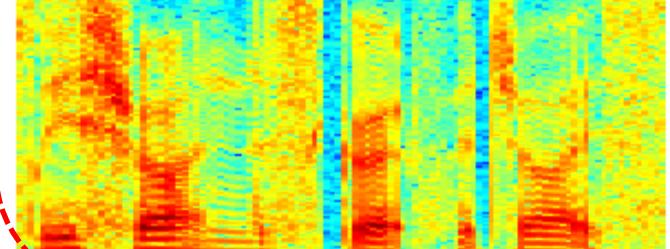
# Prosody



# Shared domain of representation

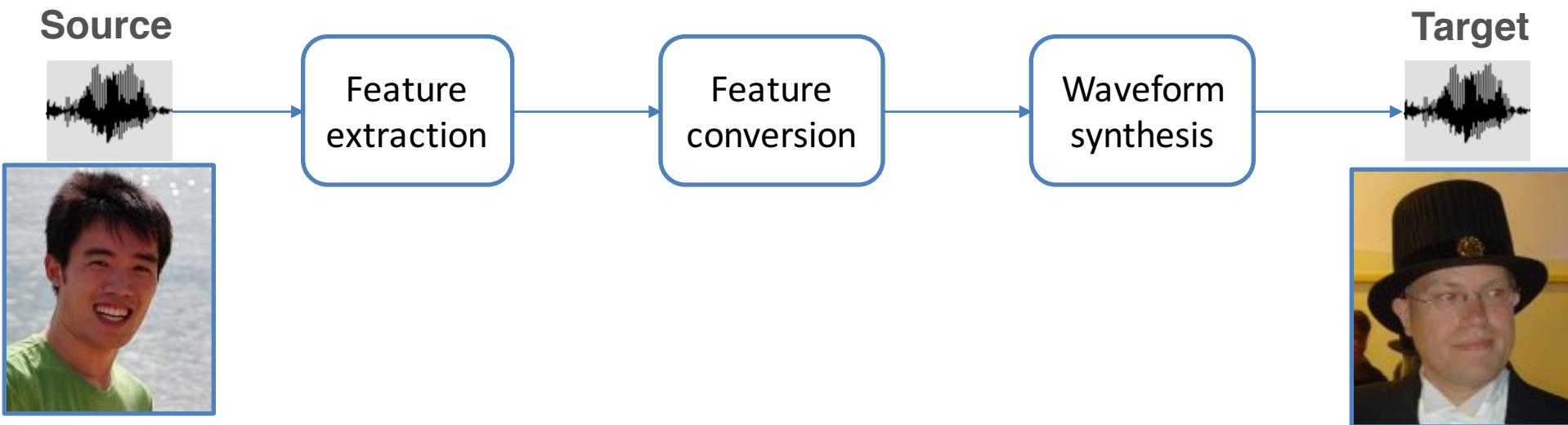
- Text-to-speech
- Voice conversion
- Automatic speaker verification

# Timbre

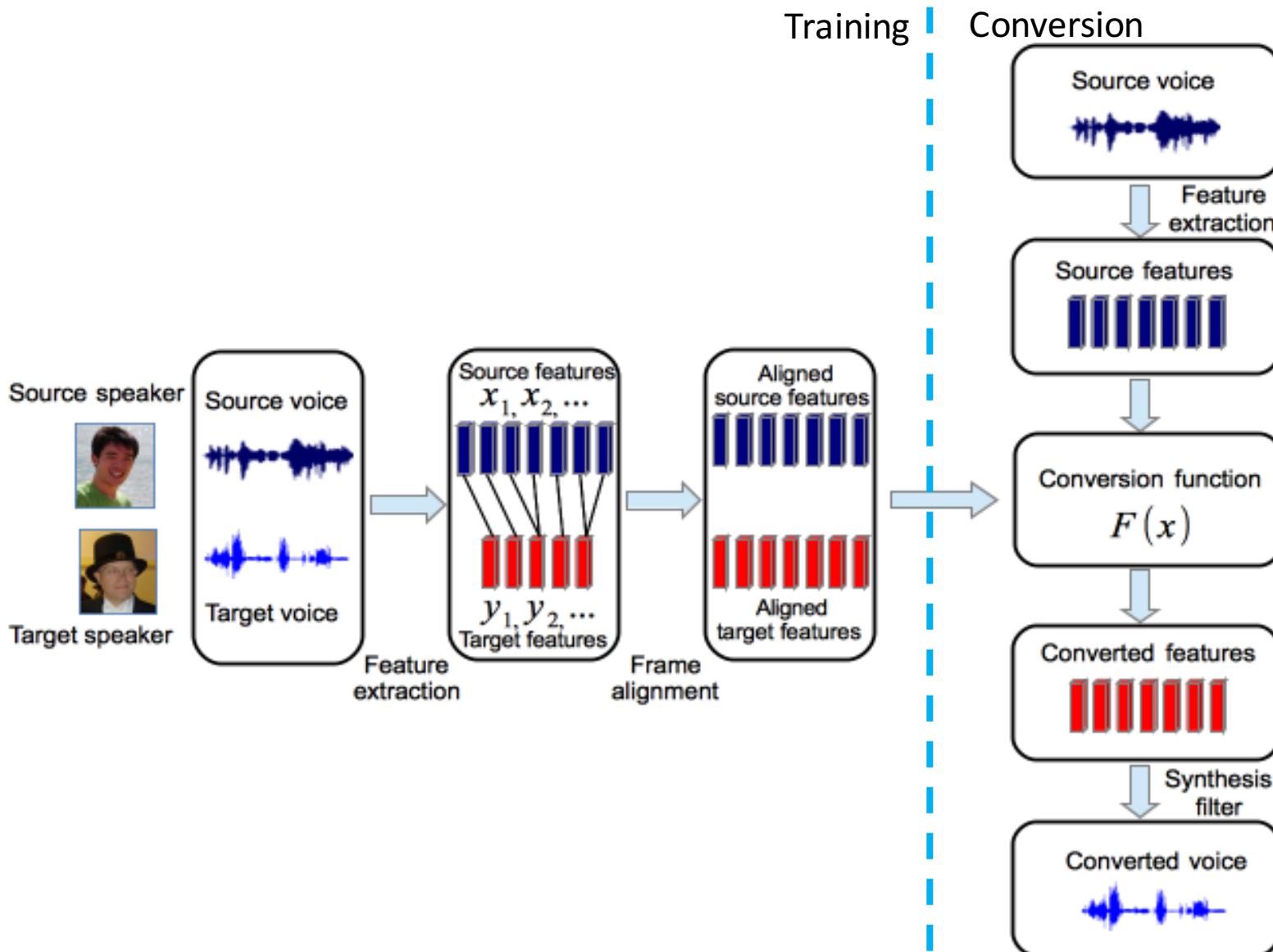


# How to convert a voice?

- Waveform to waveform conversion

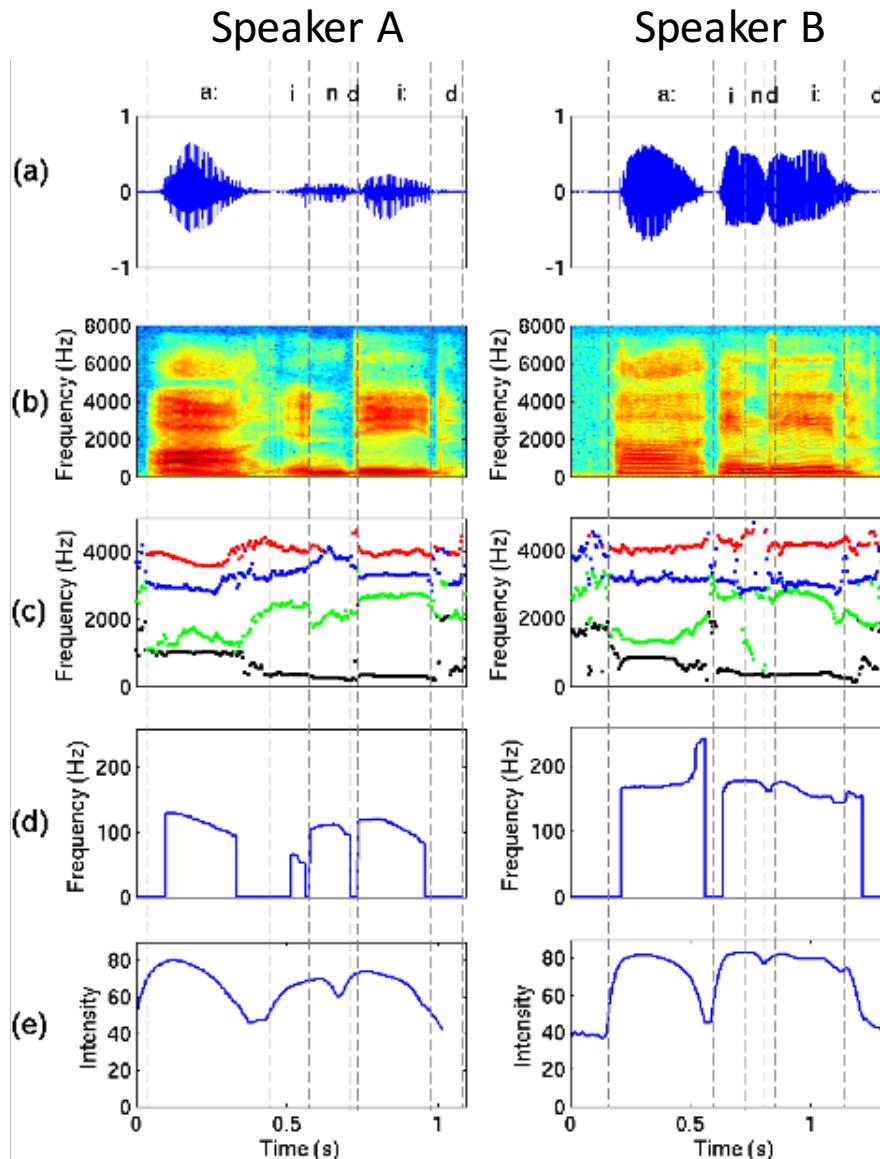


# Typical framework



# Features

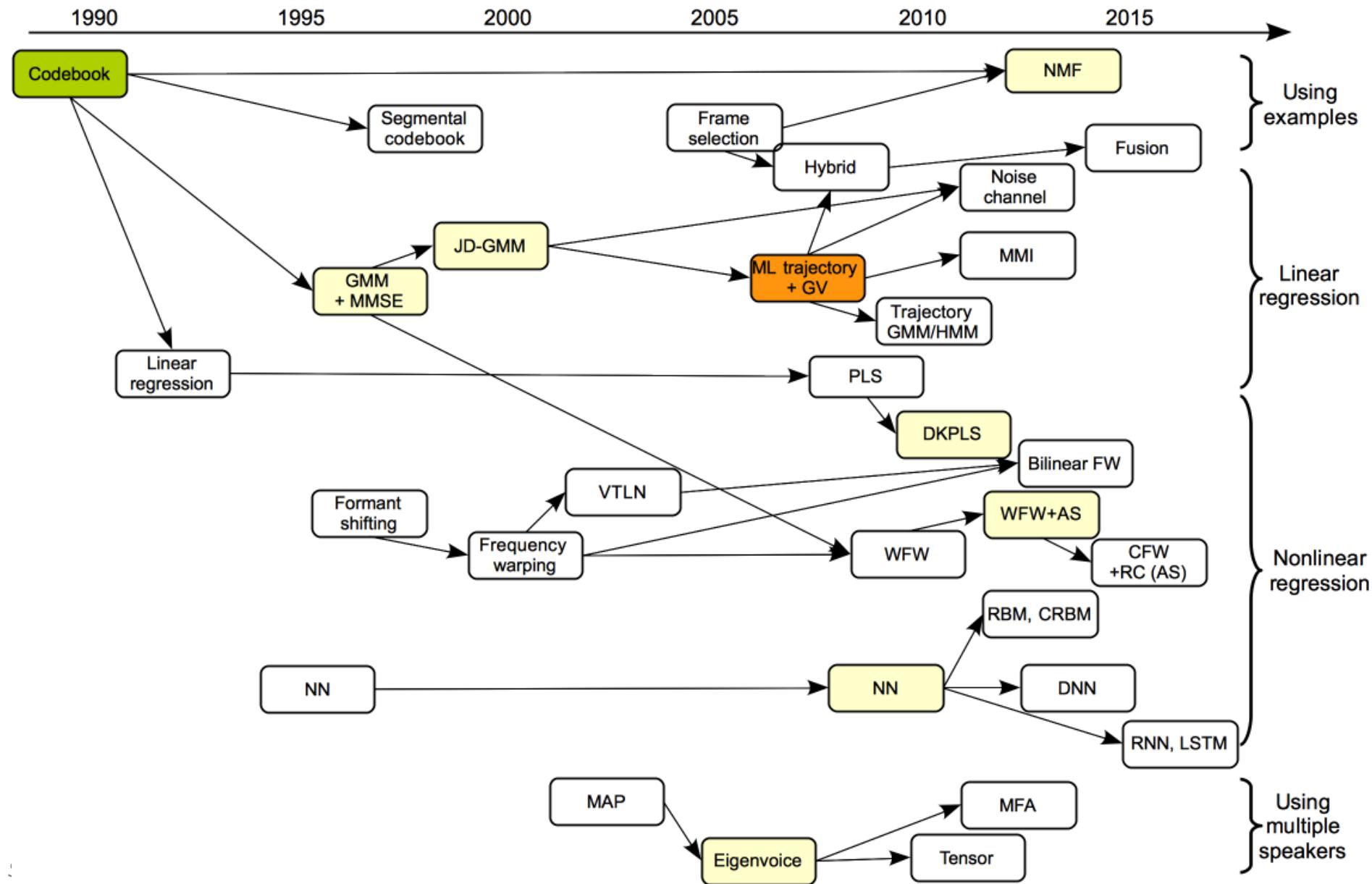
Waveform



Fundamental  
frequency

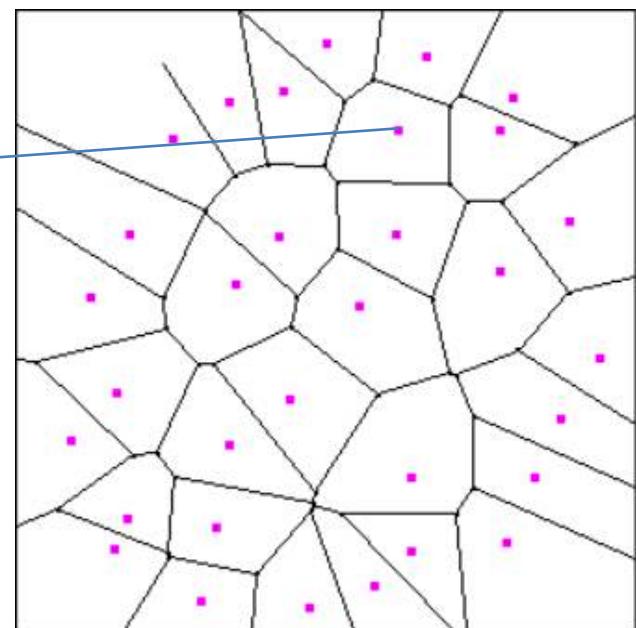
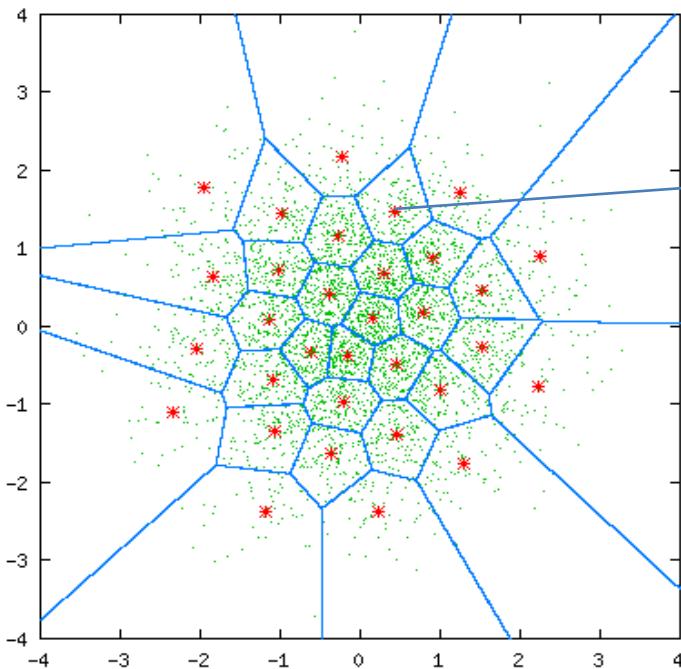
Intensity

# Progress of voice conversion



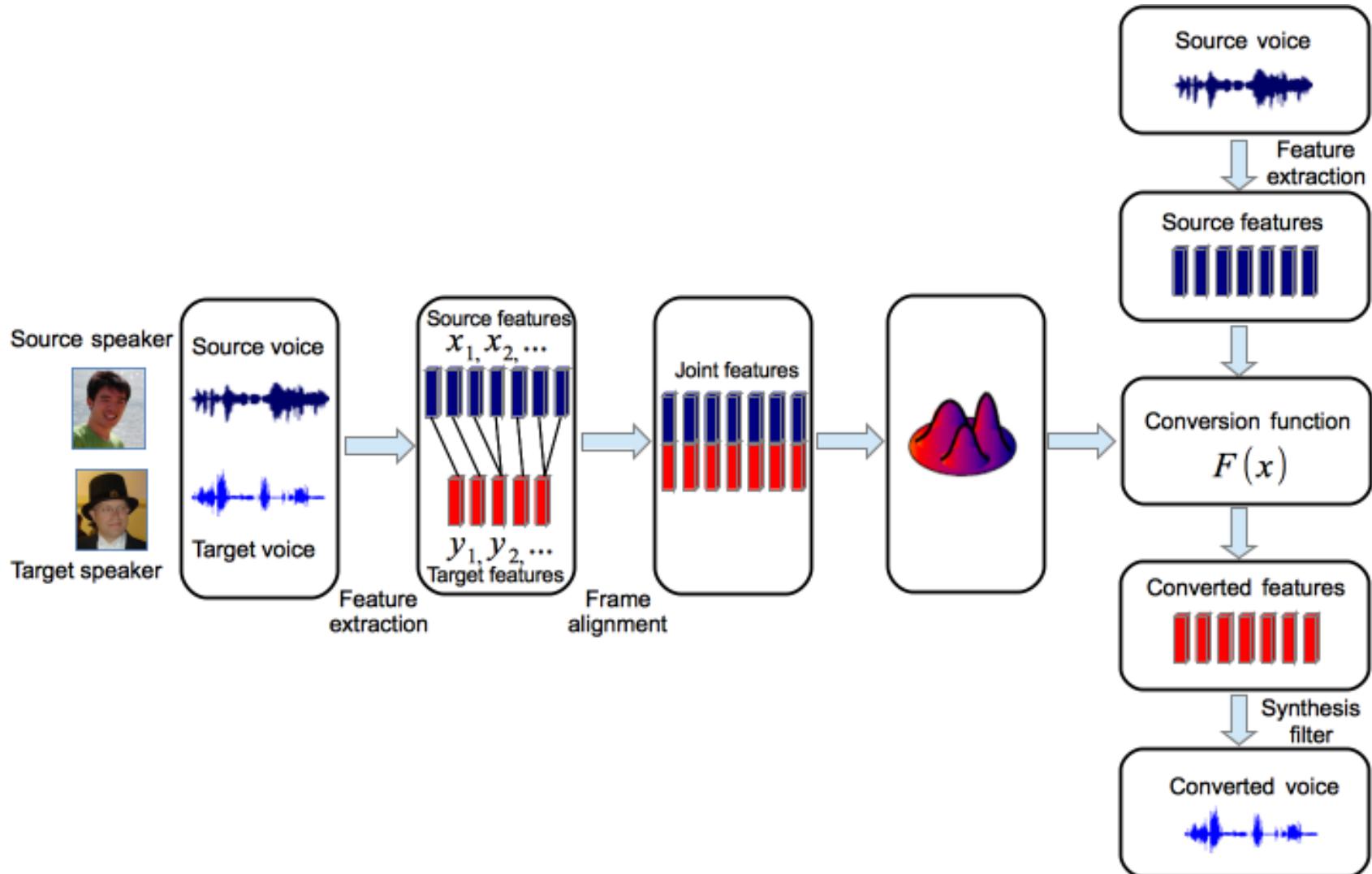
# Codebook mapping

- Vector quantisation (VQ)



Abe, Masanobu, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. "Voice conversion through vector quantization." ICASSP 1988

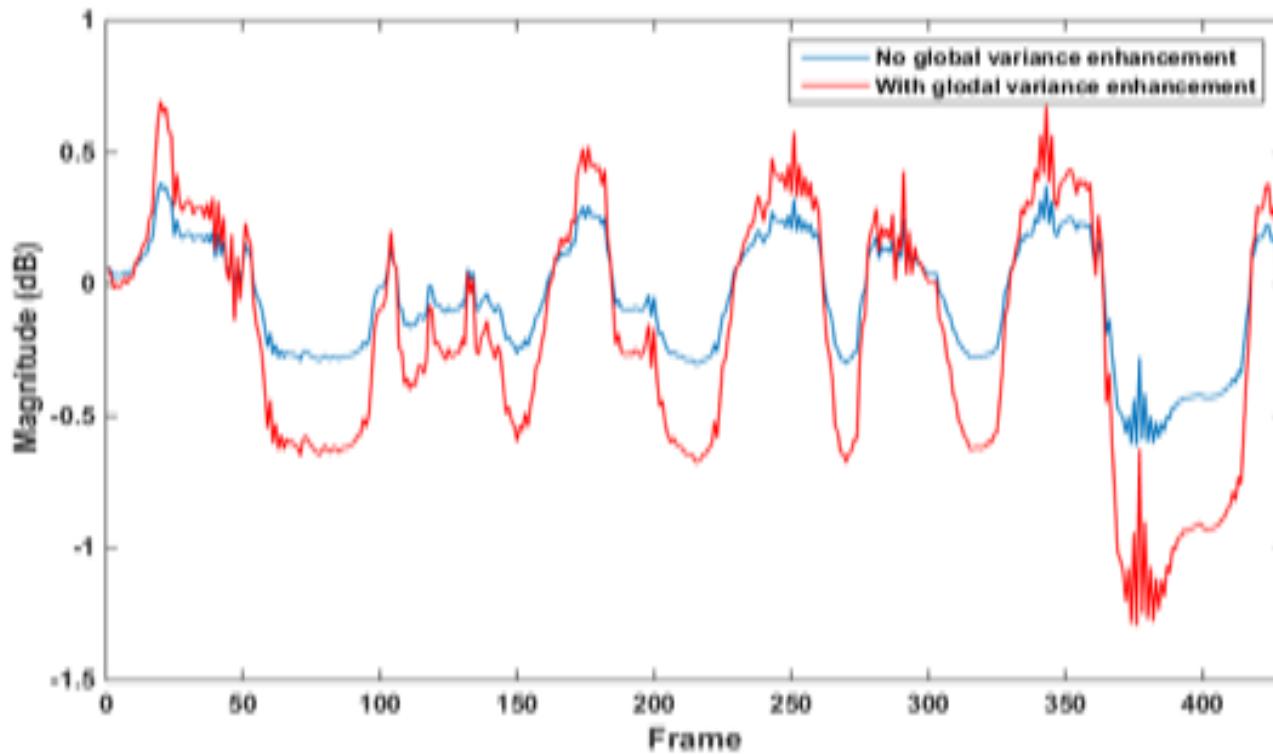
# Joint-density GMM



Alexander Kain, and Michael W. Macon. "Spectral voice conversion for text-to-speech synthesis." ICASSP 1998

# Trajectory GMM with GV

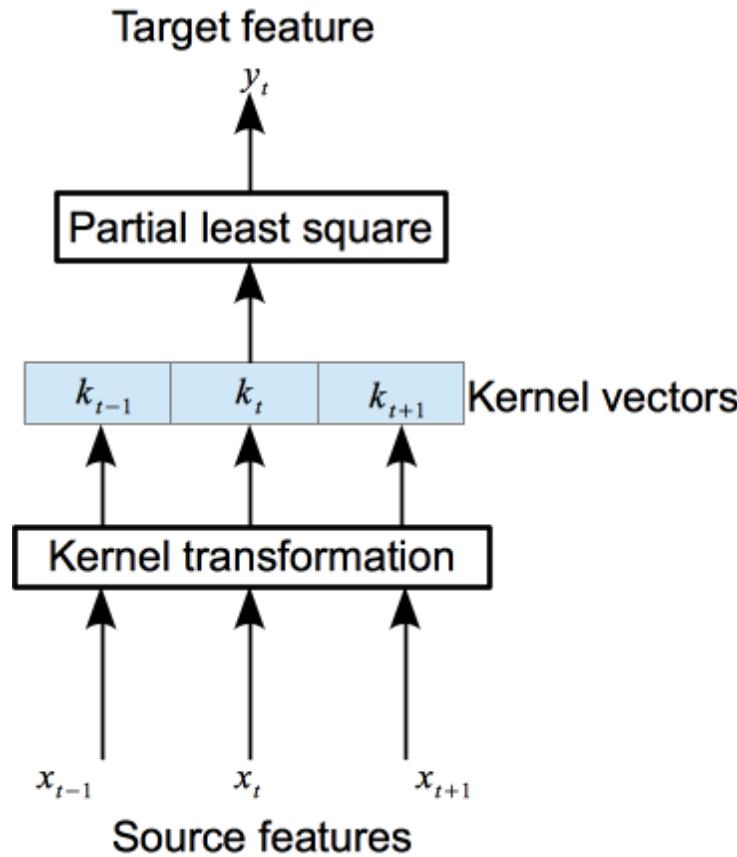
- Smooth a trajectory based on dynamic constraints
  - Same technique as that for HMM-based synthesis
- Enhance trajectory variations/dynamics



Tomoki Toda, Alan W. Black, and Keiichi Tokuda. "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory." *IEEE Transactions on Audio, Speech, and Language Processing*, 15, no. 8 (2007): 2222-2235

# Nonlinear regression

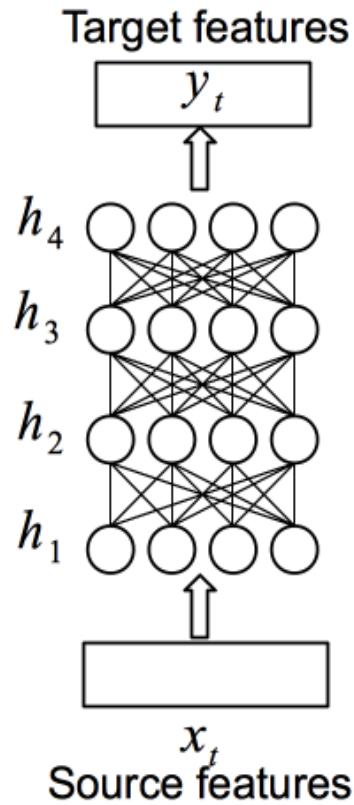
- Dynamic kernel partial least square regression (KPLS)



Helander, Elina, Hanna Silén, Tuomas Virtanen, and Moncef Gabbouj. "Voice conversion using dynamic kernel partial least squares regression." *IEEE Transactions on Audio, Speech, and Language Processing*, 20, no. 3 (2012): 806-817.

# Neural network-based VC

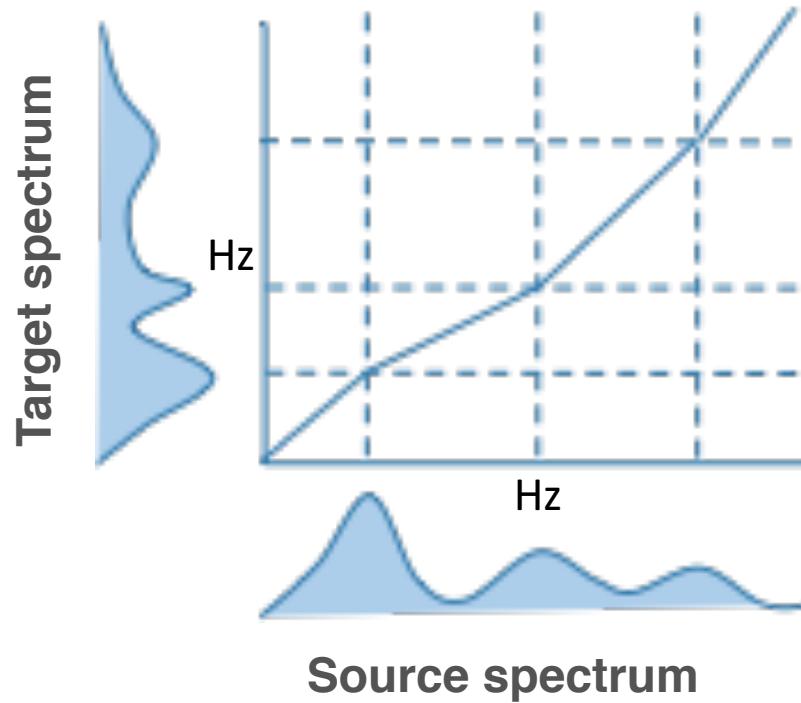
Map source features to target space by deep and/or recurrent neural networks



Srinivas Desai, Alan W. Black, B. Yegnanarayana, and Kishore Prahallad. "Spectral mapping using artificial neural networks for voice conversion." IEEE Transactions on Audio, Speech, and Language Processing, 18, no. 5 (2010): 954-964.

# Frequency warping

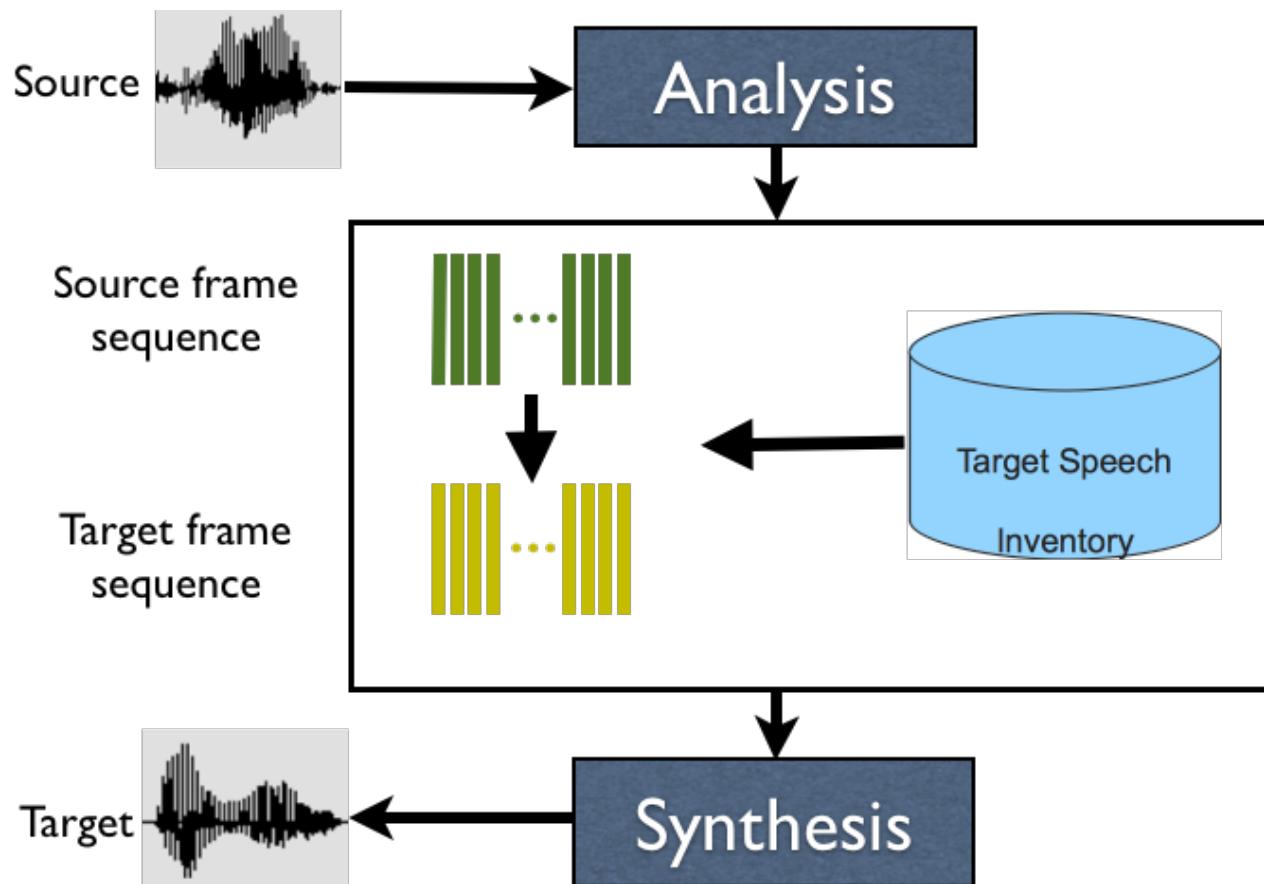
- Shifting frequency axes



Daniel Erro, Asunción Moreno, and Antonio Bonafonte. "Voice conversion based on weighted frequency warping." IEEE Transactions on Audio, Speech, and Language Processing, 18, no. 5 (2010): 922-931.

Xiaohai Tian, Zhizheng Wu, Siu Wa Lee, Nguyen Quy Hy, Eng Siong Chng, Minghui Dong, "Sparse representation for frequency warping based voice conversion", ICASSP 2015

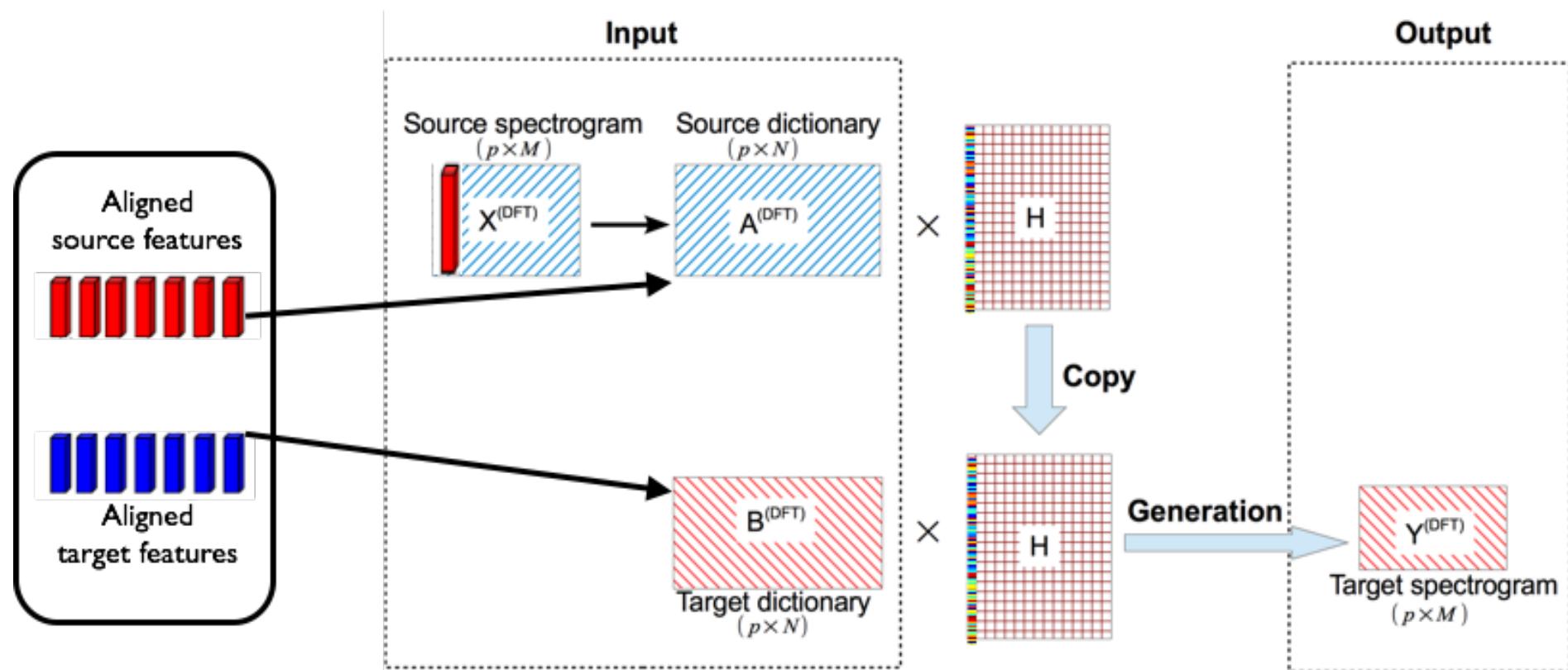
# Unit-selection based VC



Thierry Dutoit, Andre Holzapfel, Matthieu Jottrand, Alexis Moinet, J. M. Perez, and Yannis Stylianou. "Towards a voice conversion system based on frame selection." ICASSP 2007.

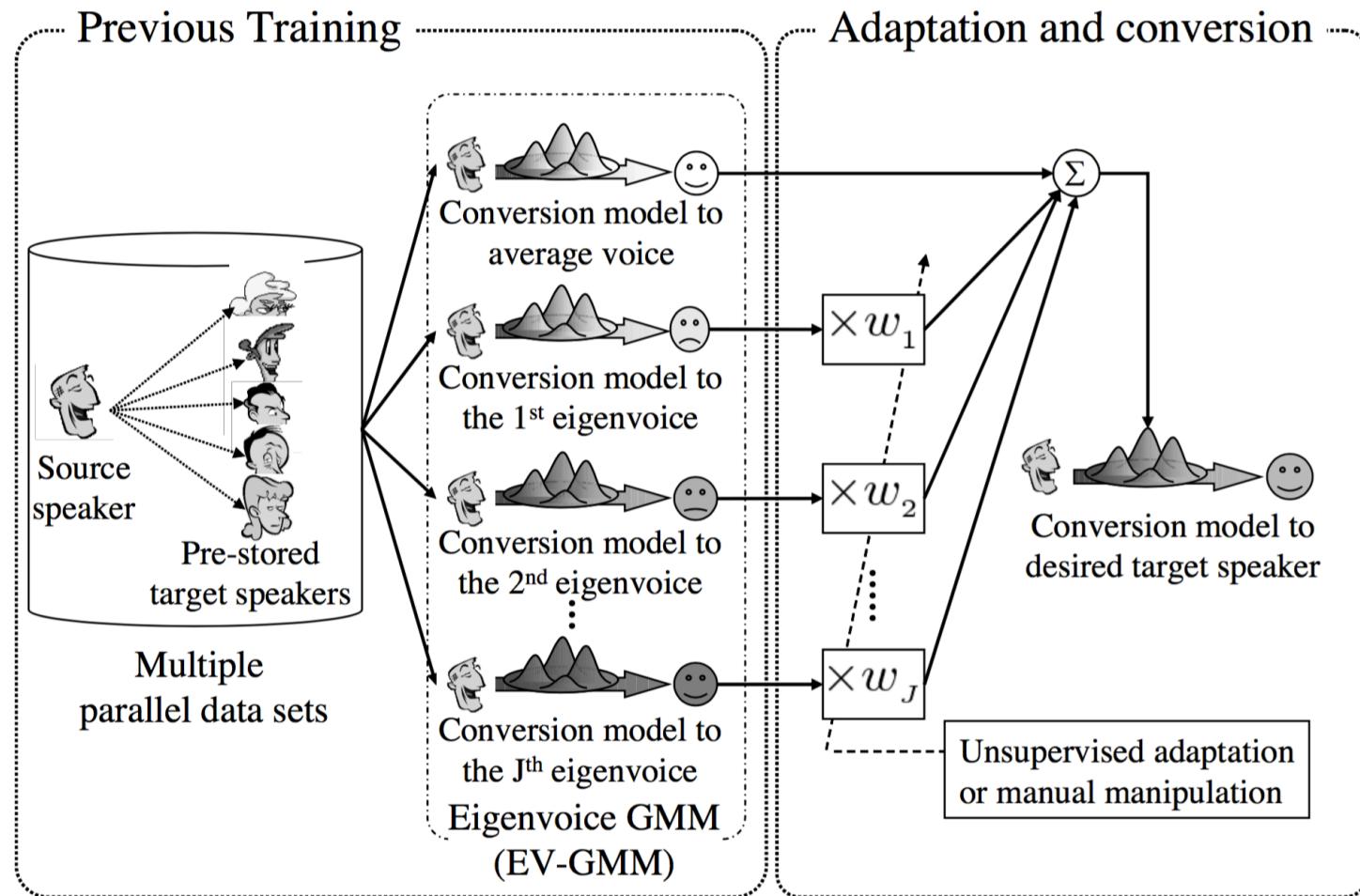
Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, "Exemplar-based unit selection for voice conversion utilizing temporal information", Interspeech 2013

# Exemplar(NMF)-based VC



Zhiheng Wu, Tuomas Virtanen, Eng Siong Chng, Haizhou Li, "Exemplar-based sparse representation with residual compensation for voice conversion", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol 22, Issue 10, pp. 1506-1521, 2014

# Eigenvoice-based voice conversion



Yamato Ohtani. "Techniques for improving voice conversion based on eigenvoices." PhD Thesis, Nara Institute of Science and Technology, 2010.  
Tomoki Toda, Yamato Ohtani, and Kiyohiro Shikano. "One-to-many and many-to-one voice conversion based on eigenvoices." ICASSP 2007.

# Tools and corpora

- Festvox: <http://festvox.org/>
  - Including GMM-based conversion with global variance enhancement
- SPTK: <http://sp-tk.sourceforge.net/>
  - Joint-density GMM conversion tools
  - Speech processing tools
- Corpora
  - CMU ARCTIC: [http://festvox.org/cmu\\_arctic/](http://festvox.org/cmu_arctic/)
  - VOICES: <https://catalog.ldc.upenn.edu/LDC2006S01>
  - VCTK: <http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html>
  - DAPS: [https://archive.org/details/daps\\_dataset](https://archive.org/details/daps_dataset)

# Voice conversion challenge 2016

- Compare and understand VC systems and approaches using a common corpus and the same protocol
- A special session at INTERSPEECH 2016
- <http://vc-challenge.org/>

# Outline

## Part 1

1. Introduction
2. Speaker verification
3. Speech synthesis
4. Voice conversion
5. Q&A

## Part 2

6. Spoofing
7. Countermeasures
8. ASVspoof 2015
9. Future
10. Q&A





# Outline

## Part 1

1. Introduction
2. Speaker verification
3. Speech synthesis
4. Voice conversion
5. Q&A

## Part 2

6. **Spoofing**
7. **Countermeasures**
8. ASVspoof 2015
9. Future
10. Q&A



# 6 & 7: Spoofing and Countermeasures

Introduction

Impersonation

Replay

Speech synthesis

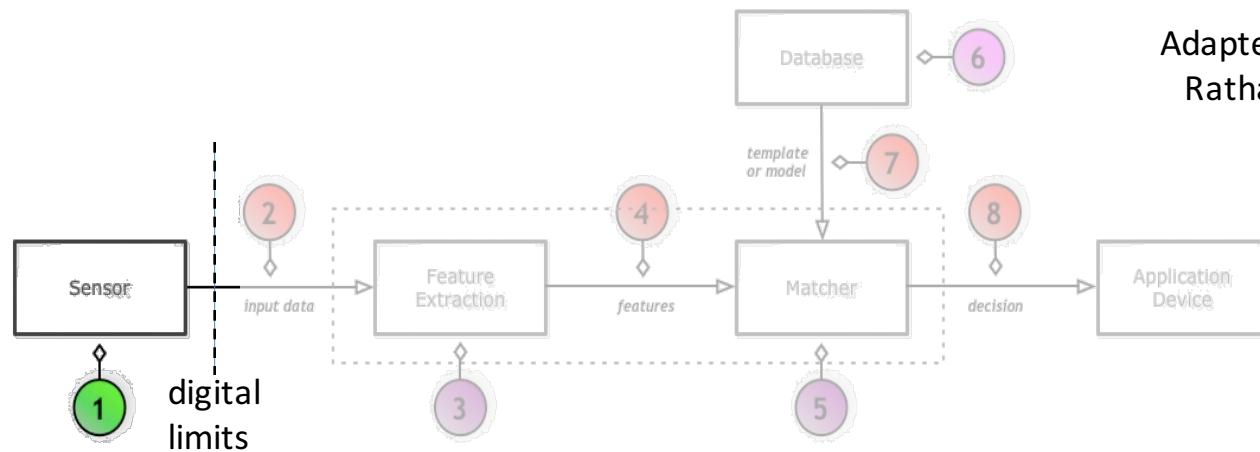
Voice conversion

Limitations

Initiatives

# Spoofing

- a.k.a. presentation attacks (ISO / IEC)
- “*persons masquerading as others in order to gain illegitimate access to sensitive or protected resources*” [Hadid et al., IEEE SPM, 2015]



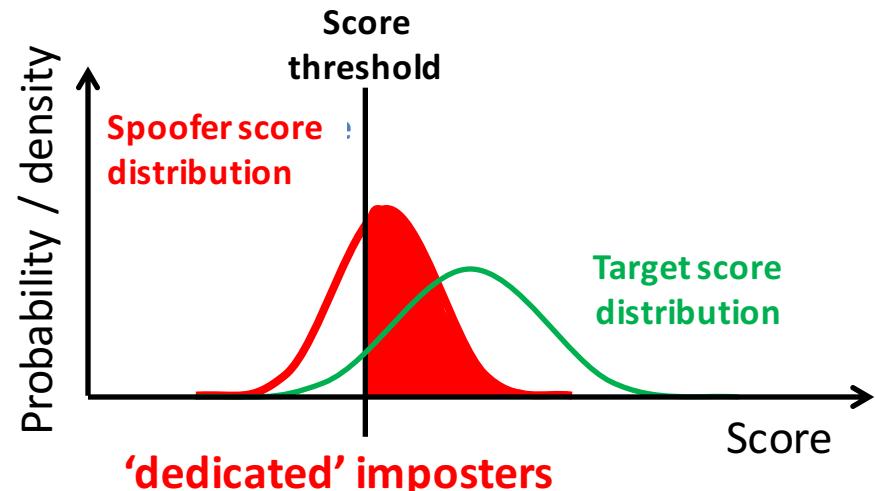
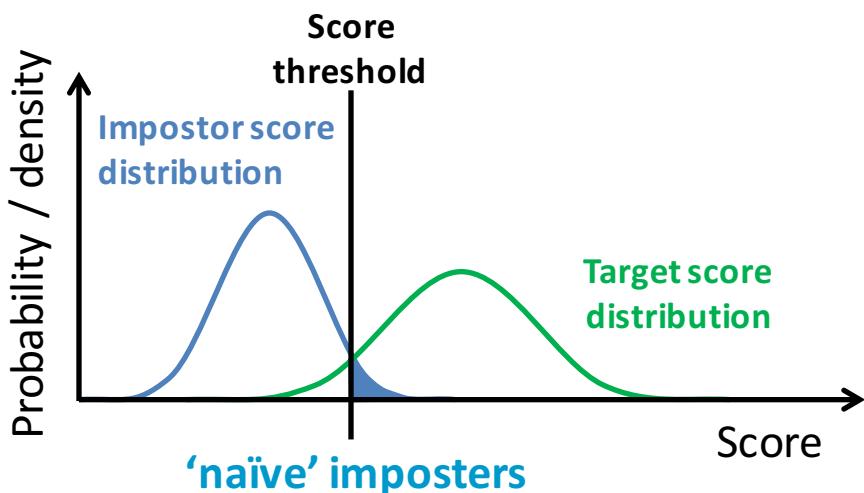
Adapted by S. Marcel from  
Ratha, Connell and Bolle,  
Proc. AVBPA, 2001

- sensor level: before and after microphone

# Spoofing

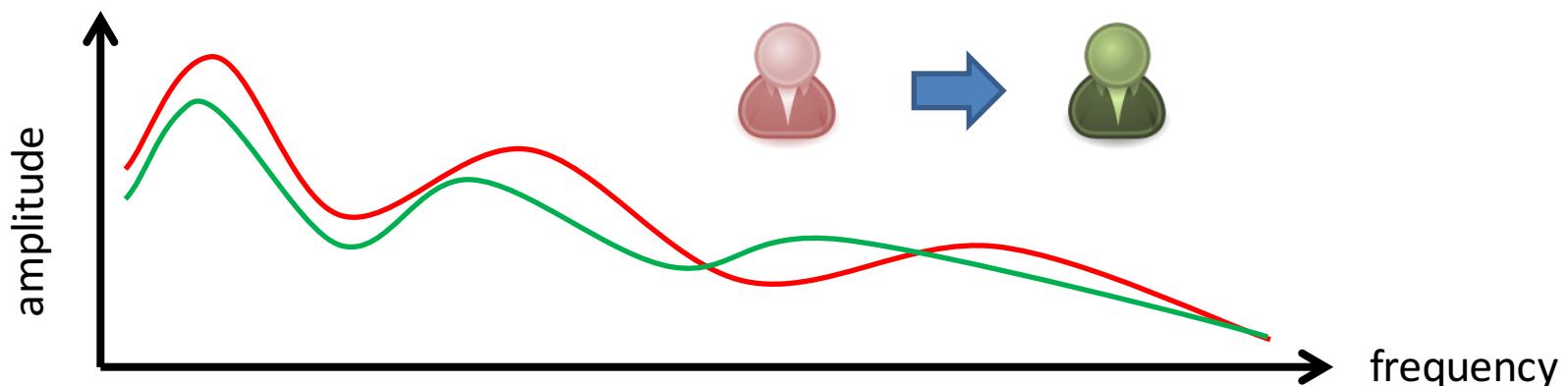
- **aim:** provoke false alarms by increasing ASV classifier score while avoiding detection

Trial	Decision	
	Accept	Reject
Target	Correct accept	<b>False reject (FR)</b>
Impostor	<b>False alarm (FA)</b>	Correct reject



# Spoofing

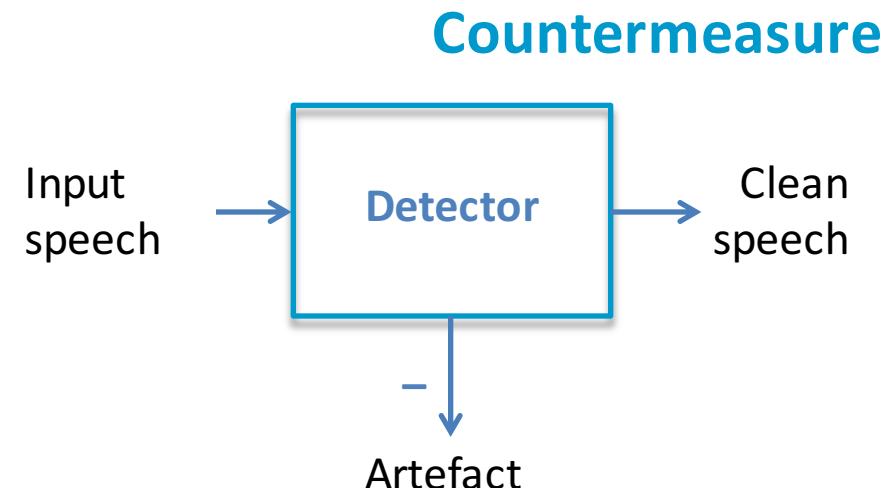
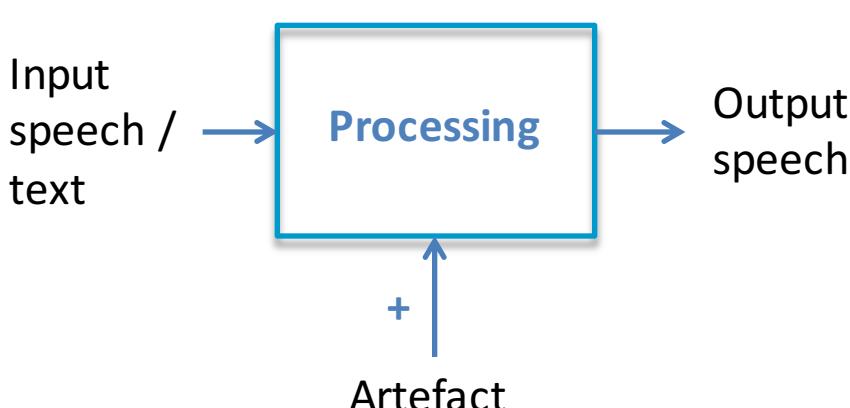
- short-term spectral estimates
- reproduce, synthesize or convert so as to resemble an enrolled, target speaker



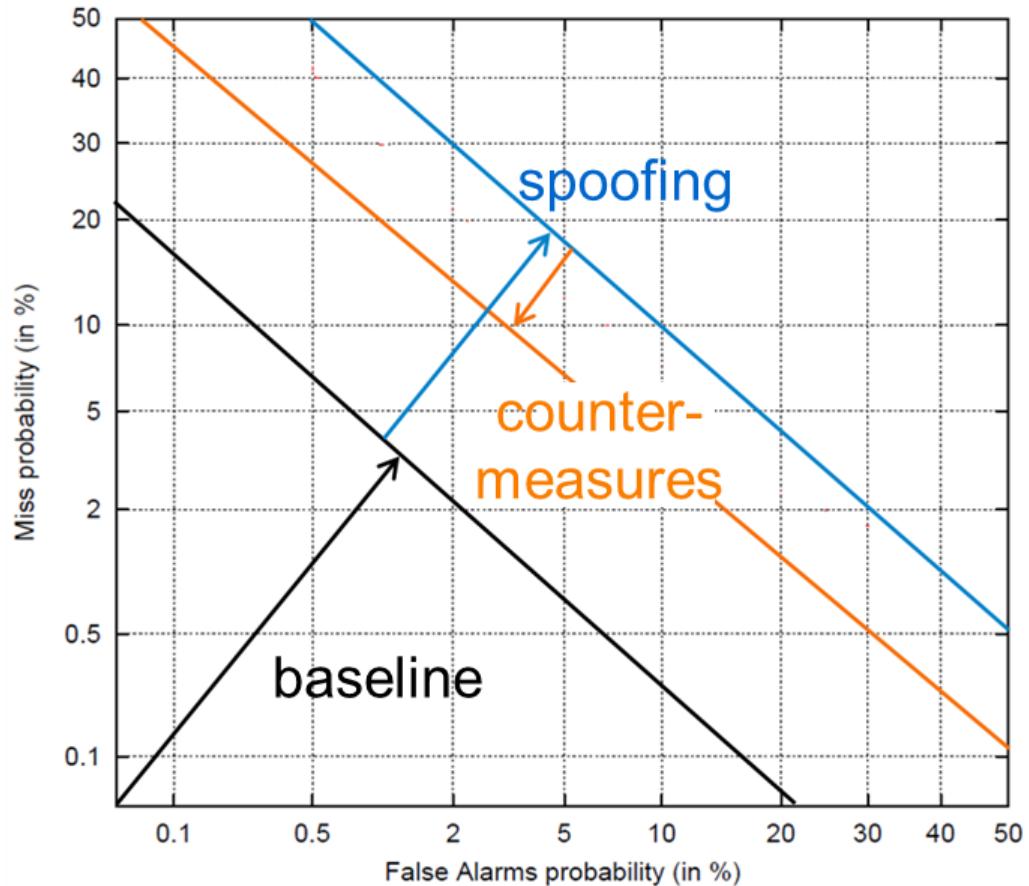
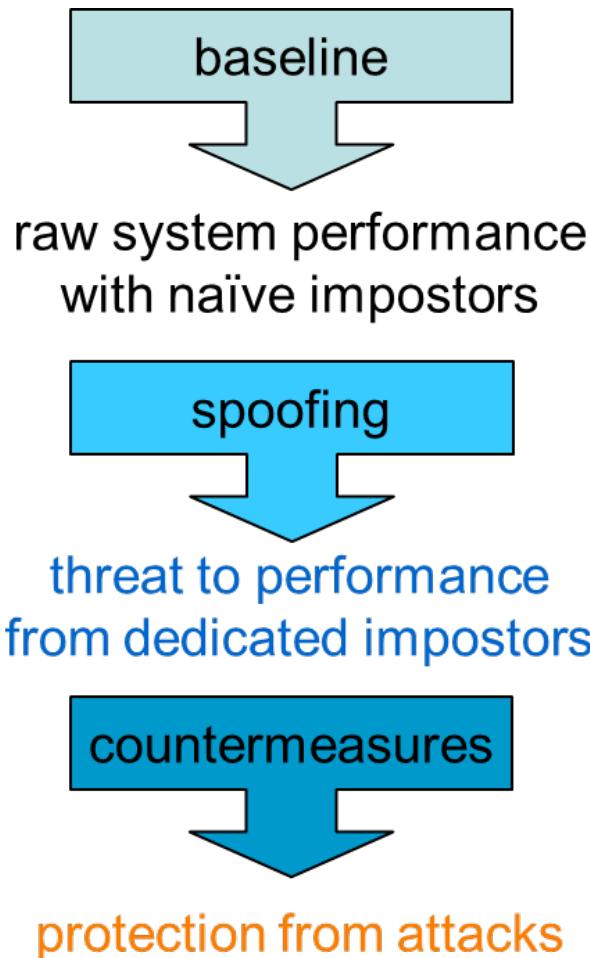
# Countermeasures

- two general approaches:
  - improve ASV robustness, i.e. feature diversity
    - implicit detection
  - dedicated countermeasures, i.e. artefact detection
    - explicit detection

## Spoofing

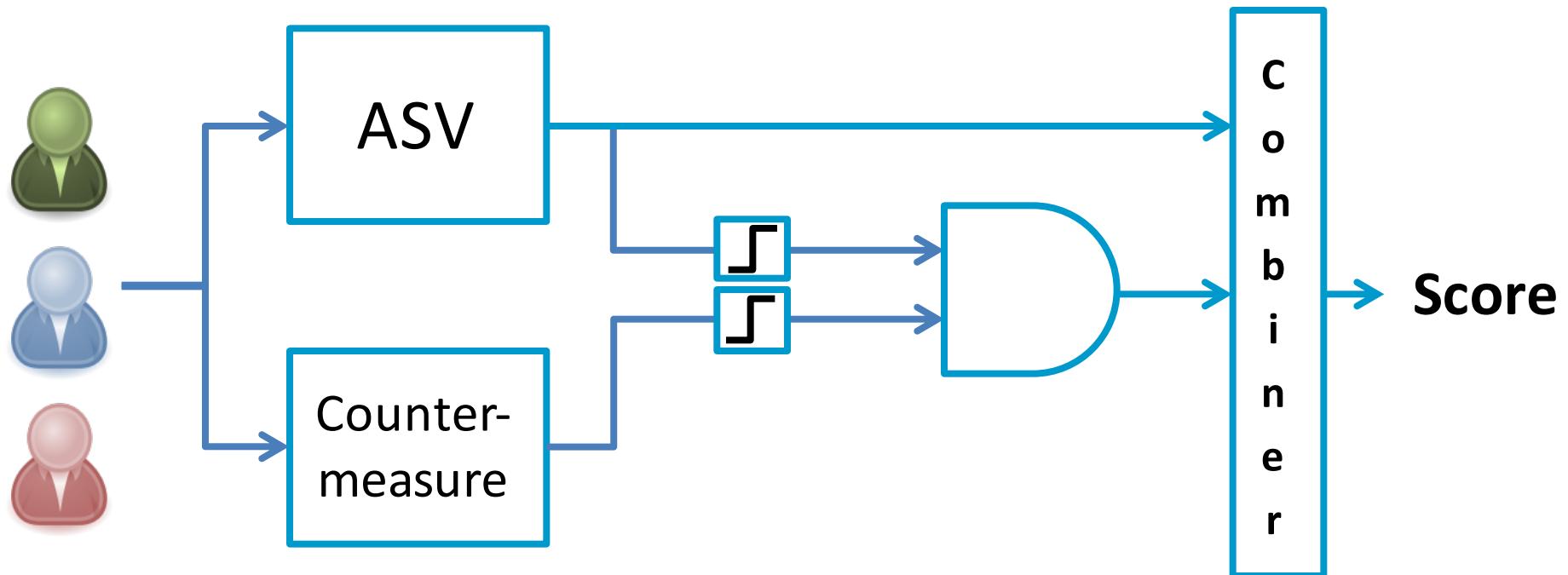


# General assessment methodology



# Integration

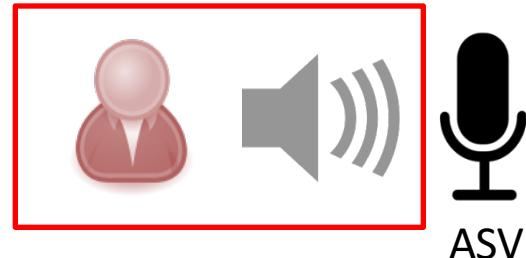
genuine clients, naïve impostors and spoofers  
detected spoofing trials set to low arbitrary score



can be preferable to assess countermeasures independently

# Impersonation

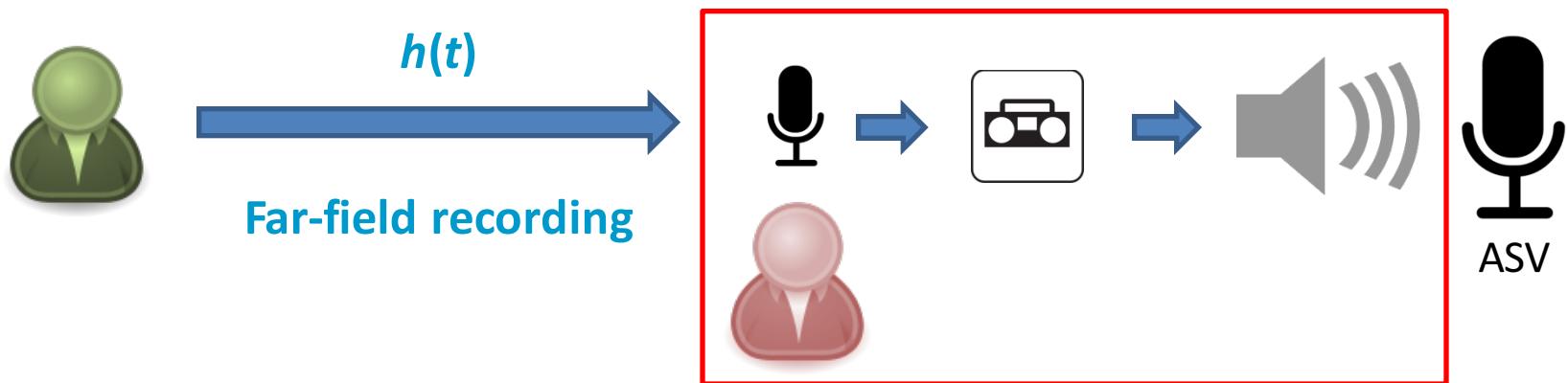
- human-altered speech to mimic timbre and prosody
- skilled attack dependent on voice similarity
- generally very few speakers
- inconsistent findings
  - human listeners v's ASV
  - prosody v's timbre



Study	# target speakers	# impersonators	ASV system	Feature	FAR or IER	
	Before spoofing	After spoofing				
Lau 2004	6	2	GMM-UBM	MFCCs	~0 %	30 ~ 35 %
Lau 2005	4	6	GMM-UBM	MFCCs	~0 %	10 ~ 60 %
Farrus 2010	5	2	k-NN	Prosodic features	5 % (IER)	22 % (IER)
Hautamäki 2013	5	1	i-vector	MFCCs	9 %	12 %

# Replay

- previously captured (concatenated) speech
- text-dependent or fixed / prompted phrase
- low-effort, low-technology attack
- generally covertly captured (e.g. passwords)



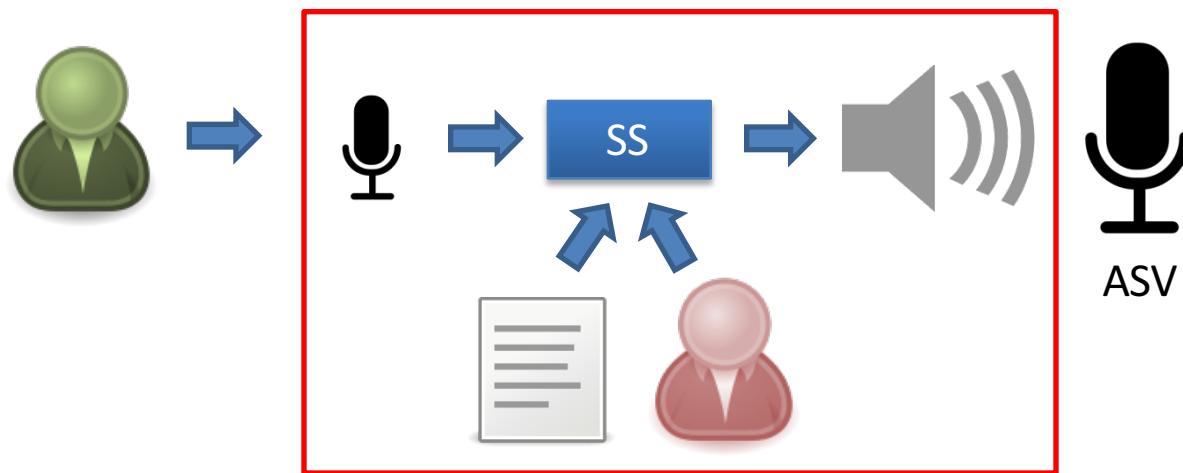
# Replay

- countermeasures:
  - audio forensic approaches, i.e. channel effects
    - e.g. sub-band ratio and modulation index [Villalba 2011]
  - passive, challenge-response
- small number of speaker, but consistent findings

Study	# target speakers	ASV system	EER/FAR	Before spoofing	After spoofing	With countermeasures	
				EER	FAR	EER	FAR
Lindberg 1999	2	Text-Dependent HMM	1 ~ 6 %	27 ~ 70 %	90 ~ 100 %	n/a	n/a
Villalba 2011	5	JFA	1%	~ 20 %	68%	0 ~ 14 %	0 ~ 17 %
Wang 2011	13	GMM-UBM	n/a	40%	n/a	10%	n/a

# Speech synthesis (1)

- artificial, synthetic speech
  - only modest requirement for target training data
- a flexible attack: no text constraints
- high-effort, high-technology, highly effective



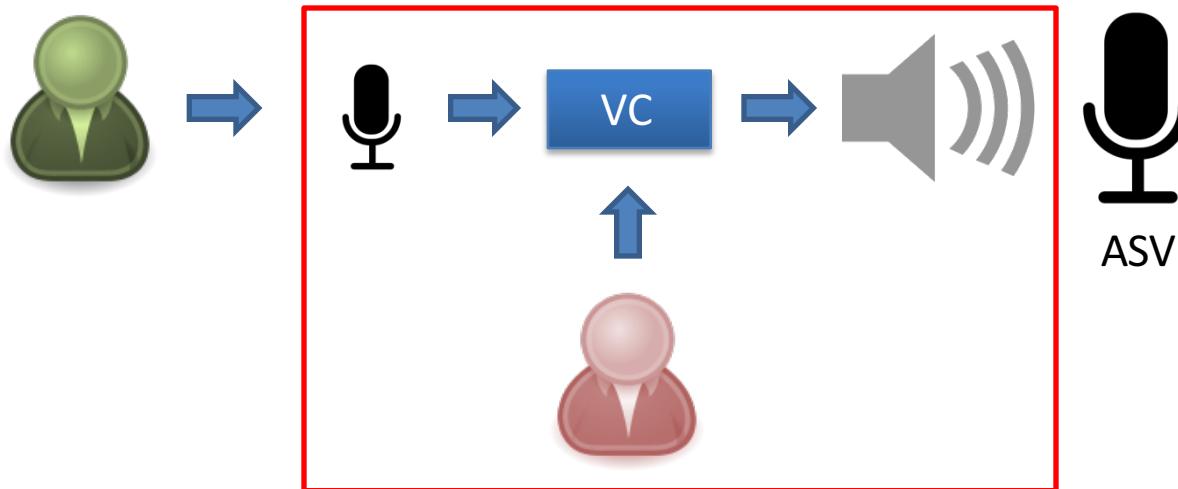
# Speech synthesis (2)

- significant studies with large, standard datasets
  - Wall Street Journal [De Leon 2012]
- universal susceptibility
- countermeasures: phase spectra and prosody

Study	# target speakers	ASV system	FAR		
			Before spoofing	After spoofing	With CMs
Lindberg 1999	2	HMM	6%	39%	n/a
Masuko 1999	20	HMM	0%	70%	n/a
De Leon 2012	283	GMM-UBM	0%	86%	2.5%
De Leon 2012	283	SVM	0%	81%	2.5%

# Voice conversion (1)

- human, converted speech
  - spectral mapping and prosody conversion
- a flexible attack: no text constraints
- potential for real-time implementations
- high-effort, high-technology, highly effective



# Voice conversion (2)

- large, standard datasets, e.g. NIST SRE
- universal susceptibility
- countermeasures: phase, prosody and dynamics

Study	# target speakers	ASV system	Before spoofing	After spoofing	With CMs
			EER/FAR	EER	FAR
Perrot 2005	n/a	GMM-UBM	~16 %	26%	~40 %
Matrouf 2006	n/a	GMM-UBM	~8 %	~63 %	~100 %
Kinnunen 2012	504	JFA	3%	8%	17%
Wu 2012b	504	PLDA	3%	11%	41%
Alegre 2013a	298	PLDA	3%	20%	~55 %
Kons 2013	750	HMM-NAP	1%	3%	36%

# Summary

Spoofing attack	Accessibility	Effectiveness (risk)		Countermeasure availability
		Text-independent	Text-dependent	
Impersonation	Low	Low/unknown	Low/unknown	Non-existant
Replay	High	Low	Low to high	Low
Speech synthesis	Medium to high	High	High	Medium
Voice conversion	Medium to high	High	High	Medium

**More objective comparisons somewhat difficult...**

# Limitations

- different datasets, protocols and metrics
  - state-of-the-art attacks
- inappropriate use of prior knowledge
  - spoofing attacks v's system
  - countermeasures v's spoofing attack
  - spoofing attacks v's countermeasure
- integration with speaker verification
- application scenario: physical / logical access
  - channel variation

**What are we doing about it?**



# TABULA RASA - EU FP7



- biometrics
  - ICAO and non-ICAO modalities
- objectives:
  - evaluate spoofing vulnerabilities
  - develop countermeasures
  - exploitation and technology transfer
  - dissemination, standards and ethics



- speaker recognition
- objectives:
  - spoofing countermeasures
  - environmental robustness
  - commercial-grade and hybrid ASV
  - scalable, trusted biometric authentication service



**Atos**

**aplcomp**



# What's missing?

- standard dataset, protocol, metric
  - a level playing field
  - advanced, state-of-the-art attacks
  - known and unknown attacks

# Outline

## Part 1

1. Introduction
2. Speaker verification
3. Speech synthesis
4. Voice conversion
5. Q&A

## Part 2

6. Spoofing
7. Countermeasures
- 8. ASVspoof 2015**
9. Future
10. Q&A



# Spoofing vs Countermeasures

shield

spear



# ASVspoof 2015

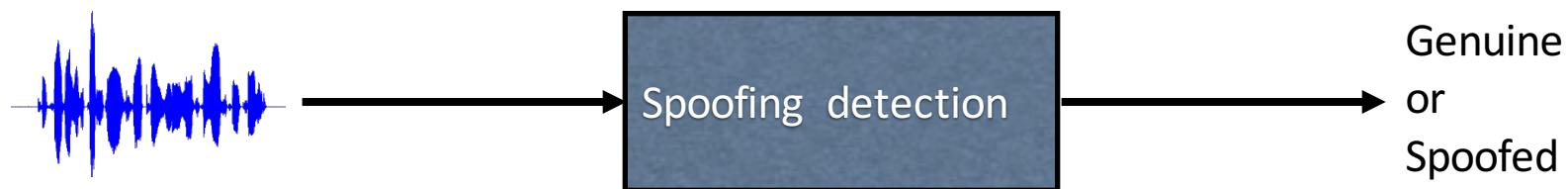
- **ASVspoof:** automatic speaker verification spoofing and countermeasures challenge
- Motivation
  - Advance the state of the art of spoofing detection
  - Standard database, common protocol, common evaluation metric

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilci, Md Sahidullah, Aleksandr Sizov, "ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge", Interspeech 2015

# The task

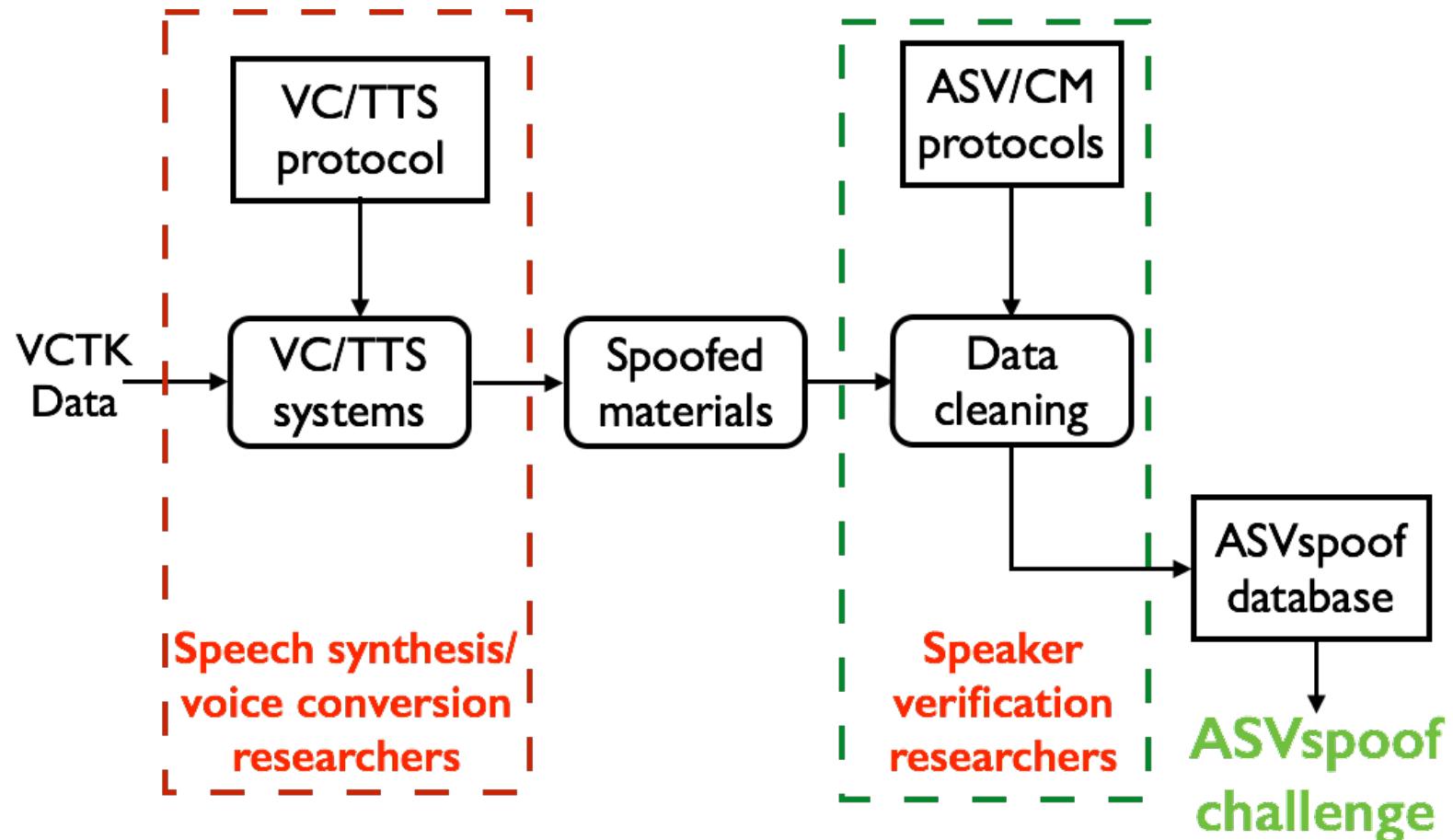
- Spoofing detection
  - To develop algorithms to discriminate between natural and spoofed speech

A speech sample



# ASVspoof database: overview

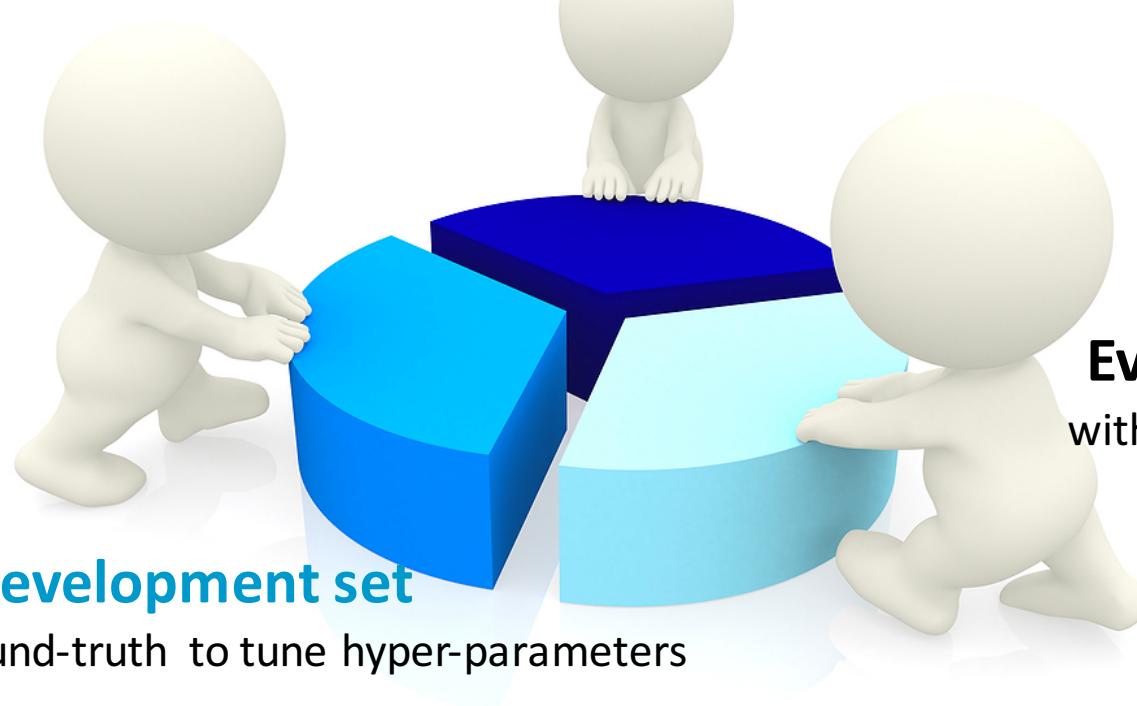
- Joint effort of speech synthesis, voice conversion and speaker verification researchers



# Database: subsets

## Training set

with known ground-truth to train or learn systems



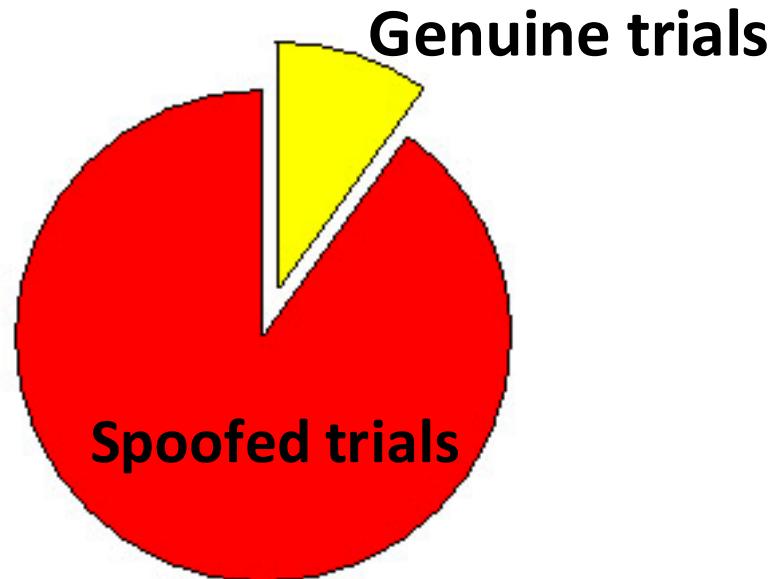
## Evaluation set

without ground-truth

## Development set

with known ground-truth to tune hyper-parameters

# Database: subsets



Clean data without channel or additive noise

# Database: subsets

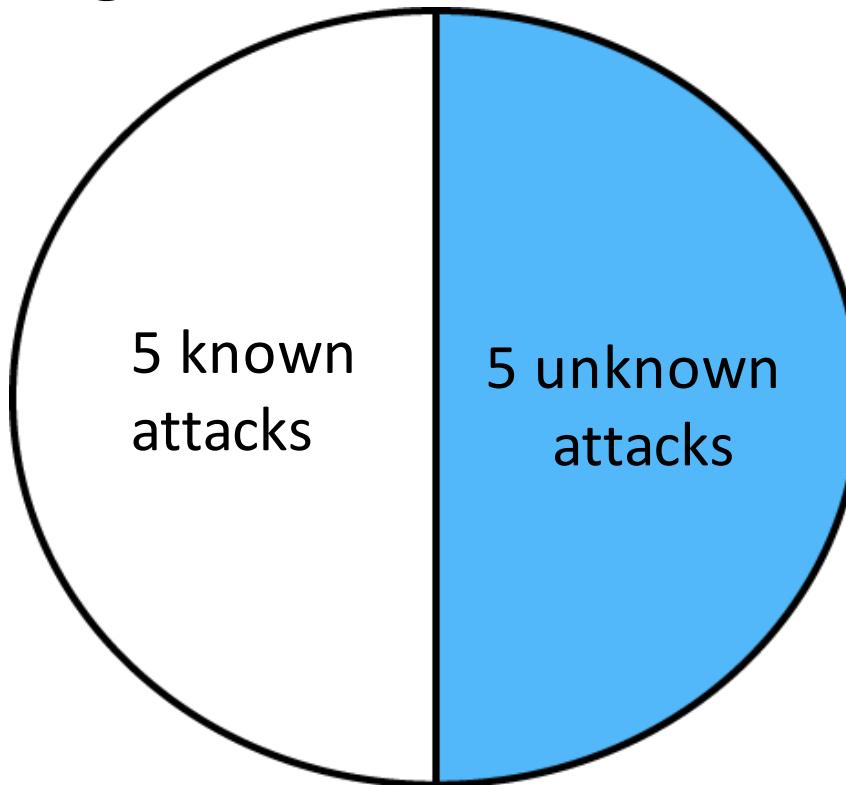
- Number of non-overlapping speakers and utterances in each subset

	# speakers		# utterances	
	Male	Female	Genuine	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	184000

To encourage gender- and speaker-independent spoofing detection

# Database: Spoofing algorithms

- 10 spoofing algorithms



Seen in **training, development**  
& **evaluation** sets

Only appear in  
**evaluation** set



# Database: known attacks

- S1 – S5: in the training, development & evaluation sets
  - S1: VC - Frame selection
  - S2: VC - Slope shifting
  - S3: TTS – HTS with 20 adaptation sentences
  - S4: TTS – HTS with 40 adaptation sentences
  - S5: VC – Festvox (<http://festvox.org/>)





# Database: unknown attacks

- S6 – S10: Only appear in the evaluation set
  - S6: VC – ML-GMM with GV enhancement
  - S7: VC – Similar to S6 but using LSP features
  - S8: VC – Tensor (eigenvoice)-based approach
  - S9: VC – Nonlinear regression (KPLS)
  - S10: TTS – MARY TTS unit selection (<http://mary.dfki.de/>)



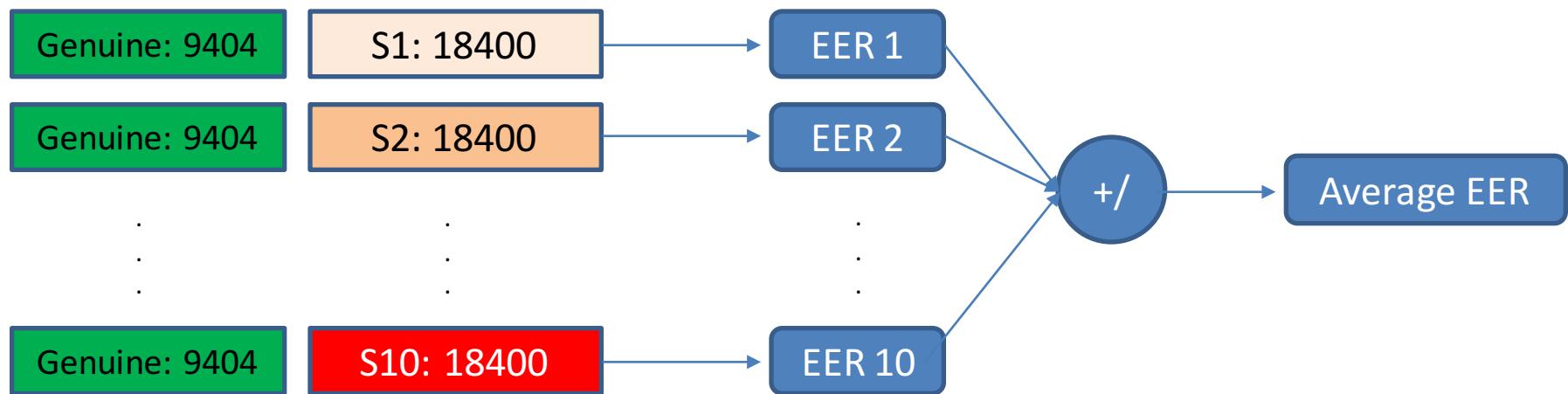
# Database: spoofing algorithms

- Summary of spoofing algorithms implemented

	# utterances			Algorithm	Vocoder
	Training	Development	Evaluation		
Genuine	3750	3497	9404	None	None
S1	2525	9975	18400	VC :Frame-selection	STRAIGHT
S2	2525	9975	18400	VC: Slope-shifting	STRAIGHT
S3	2525	9975	18400	SS: HMM	STRAIGHT
S4	2525	9975	18400	SS: HMM	STRAIGHT
S5	2525	9975	18400	VC: GMM	MLSA
S6	0	0	18400	VC: GMM	STRAIGHT
S7	0	0	18400	VC: GMM	STRAIGHT
S8	0	0	18400	VC: Tensor	STRAIGHT
S9	0	0	18400	VC: KPLS	STRAIGHT
S10	0	0	18400	SS: unit-selection	None

# Evaluation metric

- Average Equal Error Rate (EER)



# Evaluation task

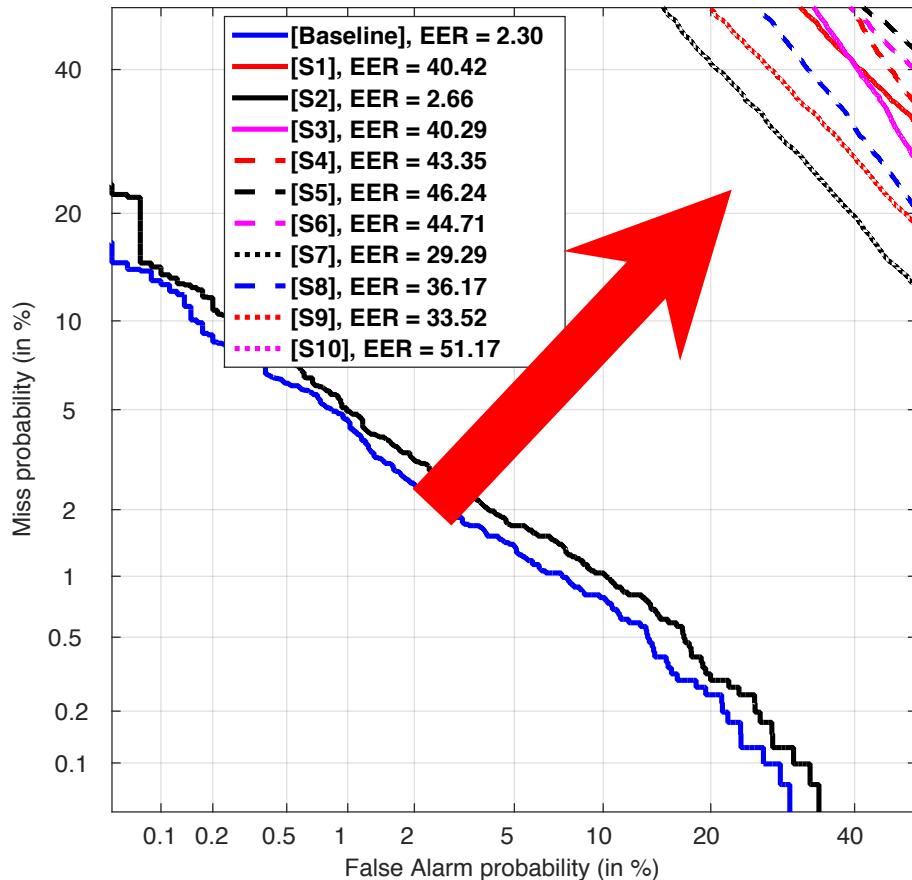
- Each participant is allowed to submit up to six systems
  - Only the primary score under the common training condition is used for ranking

Submission	Training condition	
	Common	Flexible
Primary	<b>Required</b>	Optional
Contrastive1	Optional	Optional
Contrastive2	Optional	Optional

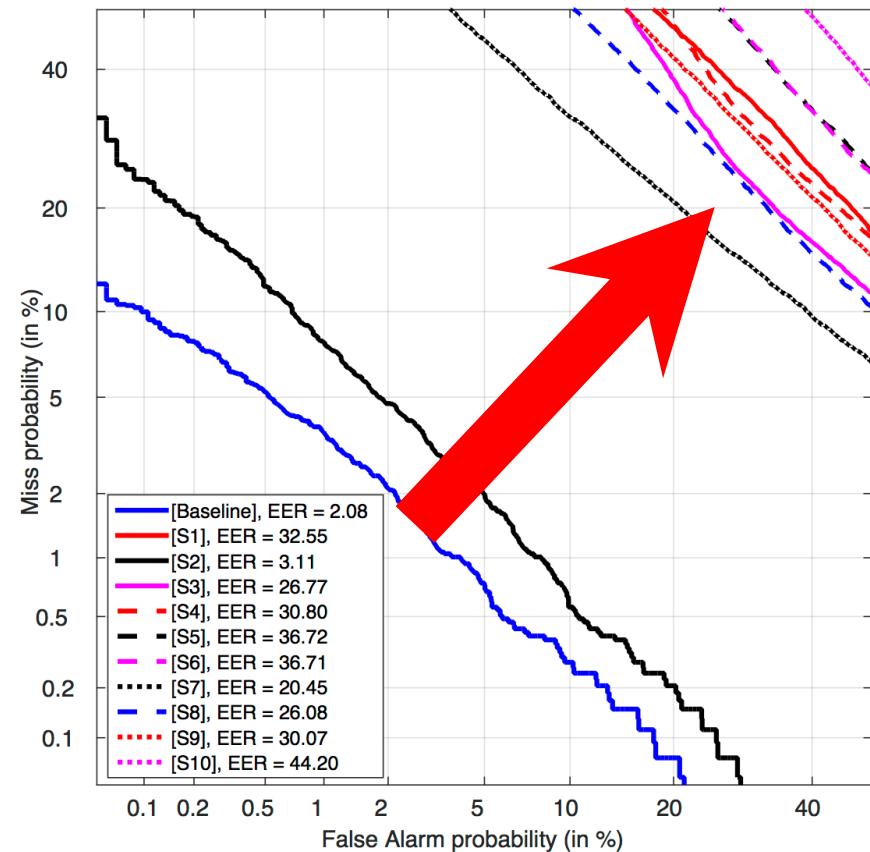
- Common condition: can only use the defined training data
- Flexible condition: can use any training data

# Speaker verification performance

- i-vector-PLDA system



Male



Female

# The challenge participation

- 28 teams from 16 countries requested the challenge database
- 16 teams submitted results by the deadline
- Received 16 primary submissions and 27 additional submissions

# Challenge results

- EERs of the primary tasks from 16 teams

Team	Known attacks (S1 - S5)	Unknown attacks (S6 - S10)	Average (all)
A	0.408	2.013	1.211
B	0.008	3.922	1.965
C	0.058	4.998	2.528
D	0.003	5.231	2.617
E	0.041	5.347	2.694
F	0.358	6.078	3.218
G	0.405	6.247	3.326
H	0.67	6.041	3.355
I	0.005	7.447	3.726
J	0.025	8.168	4.097
K	0.21	8.883	4.547
L	0.412	13.026	6.719
M	8.528	20.253	14.391
N	7.874	21.262	14.568
O	17.723	19.929	18.826
P	21.206	21.831	21.518

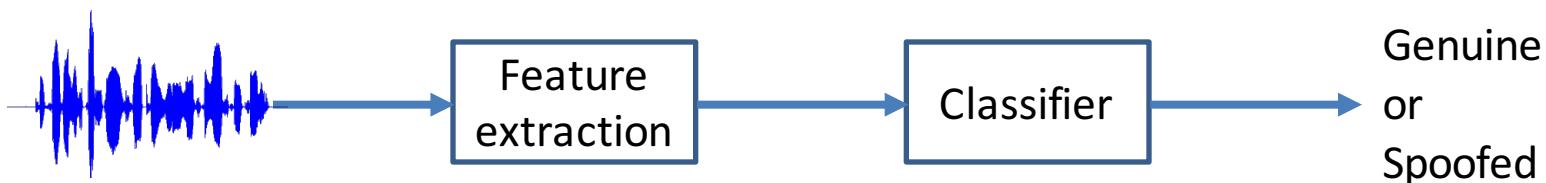
# Challenge results

- Team names

Team	Average (all)	Average (without S10)	S10	Team name
A	1.211	0.402	8.490	DA-IICT
B	1.965	0.008	19.571	STC
C	2.528	0.076	24.601	SJTU
D	2.617	0.003	26.142	NTU
E	2.694	0.060	26.393	CRIM
F	3.218	0.400	28.581	
G	3.326	0.360	30.021	
H	3.726	0.021	37.068	
I	3.898	0.703	32.651	
J	4.097	0.029	40.708	
K	4.547	0.203	43.638	
L	6.719	3.478	35.890	
M	14.391	12.482	31.574	
N	14.568	11.299	43.991	
O	18.826	16.304	41.519	
P	21.518	18.786	46.102	

# State of the art

- System A achieved the best overall performance and the best performance on S10
- System D achieved the best performance on S1-S9

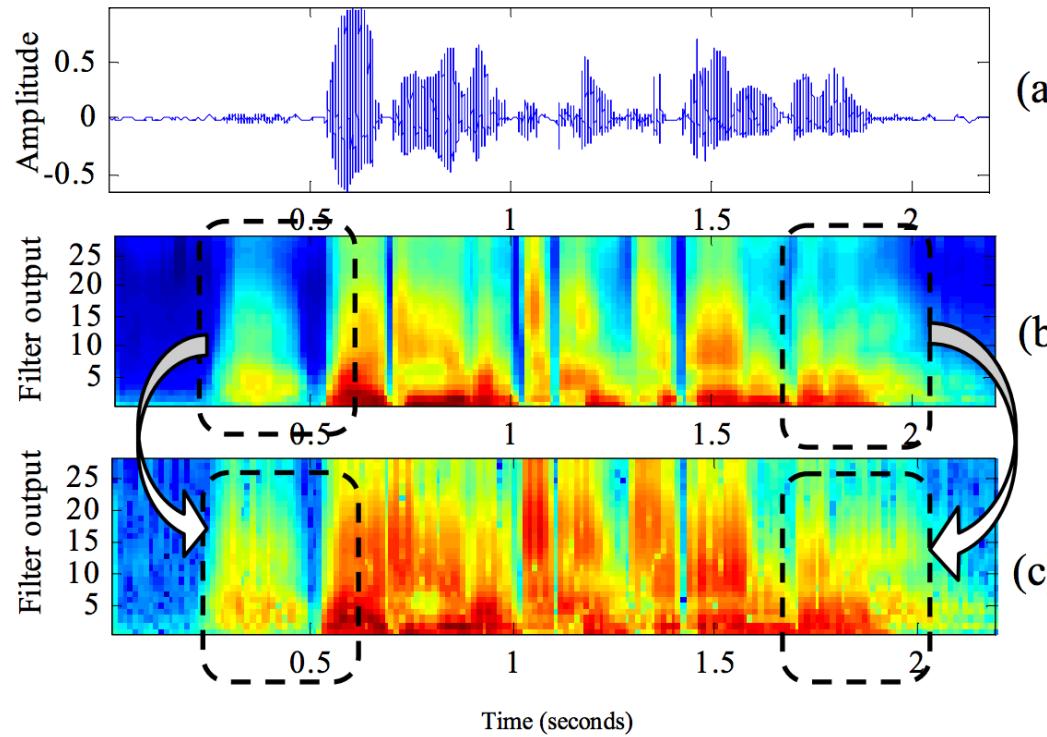


Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

# System A (DA-IICT)

Feature extraction

- CFCCIF: cochlear filter cepstral coefficients plus instantaneous frequency
- MFCC



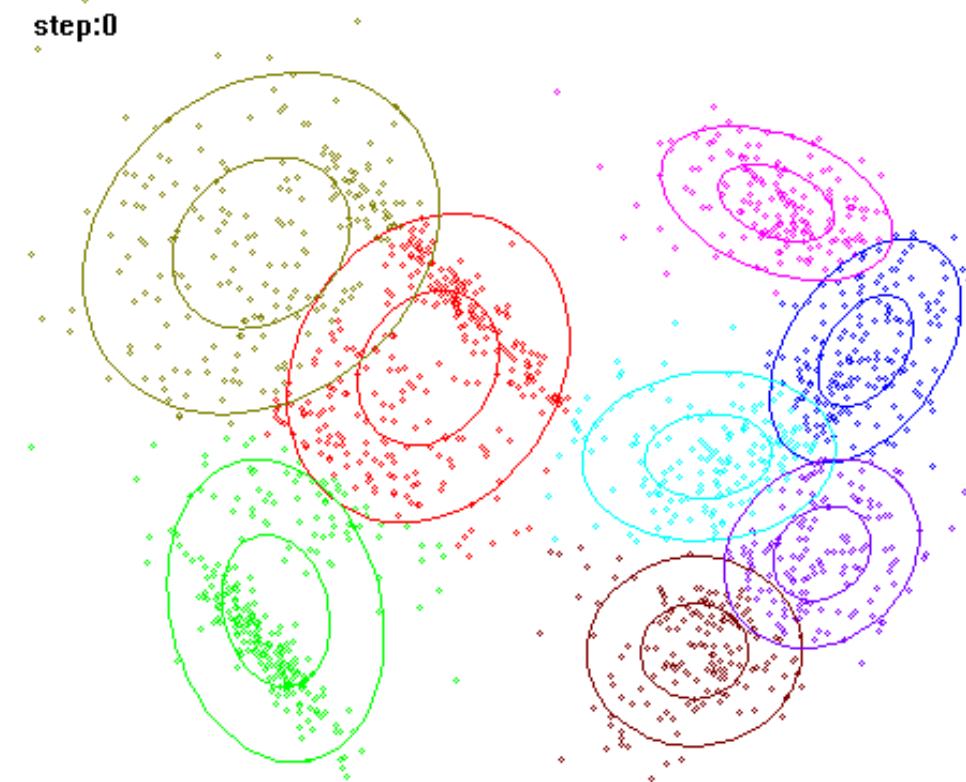
Tanvina B. Patel, Hemant A. Patil, "Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural vs. Spoofed Speech", Interspeech 2015

Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

# System A (DA-IICT)

Classifier

- GMM: log-likelihood ratio
- Score fusion

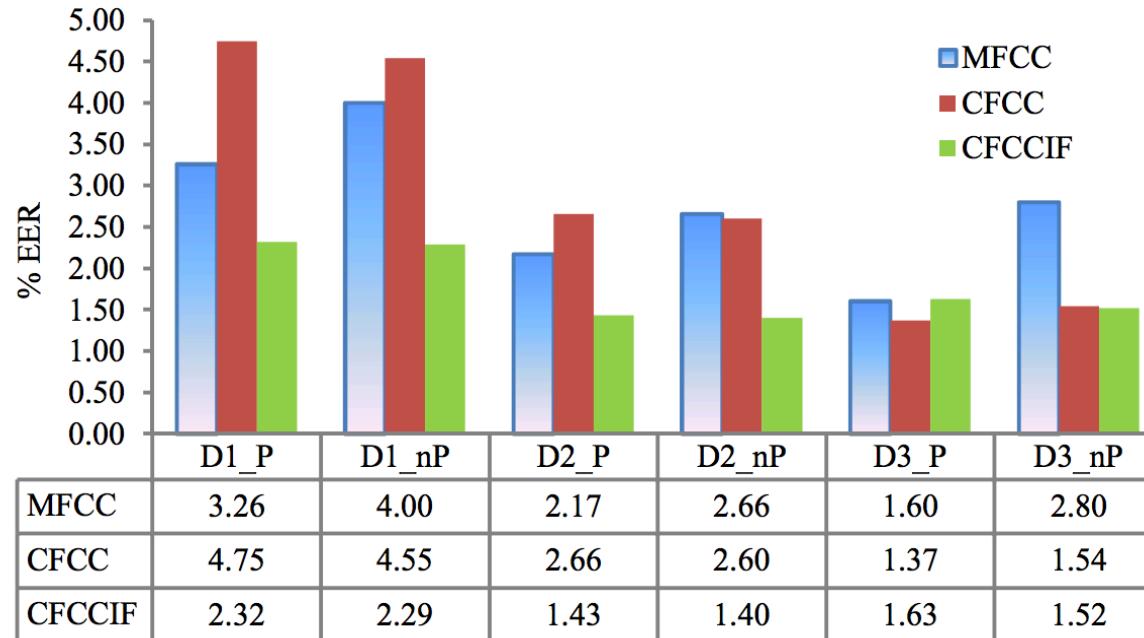


Tanvina B. Patel, Hemant A. Patil, "Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural vs. Spoofed Speech", Interspeech 2015

Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

# System A (DA-IICT): why success?

- The CFCCIF feature works well in detecting the unit selection attack, S10



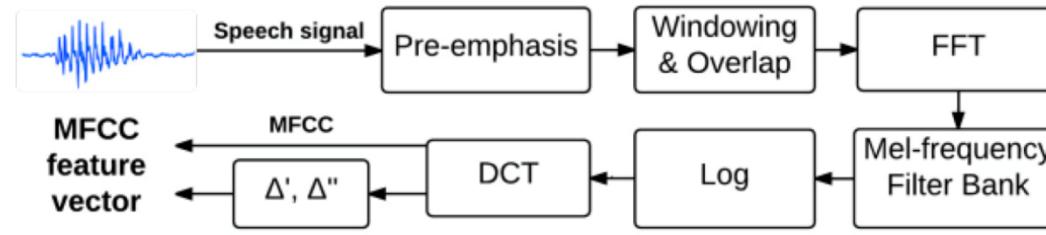
Submission	Known attacks (% EER)					Unknown attacks (% EER)					S10
	S1	S2	S3	S4	S5	S6	S7	S8	S9		
A: DA-IICT	0.1013	0.8629	0.0000	0.0000	1.0753	0.8462	0.2416	0.1417	0.3463	8.4900	
Average (Proposed)	0.407899					2.013162					
Avg. of 16 submissions	3.337					9.294					

Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

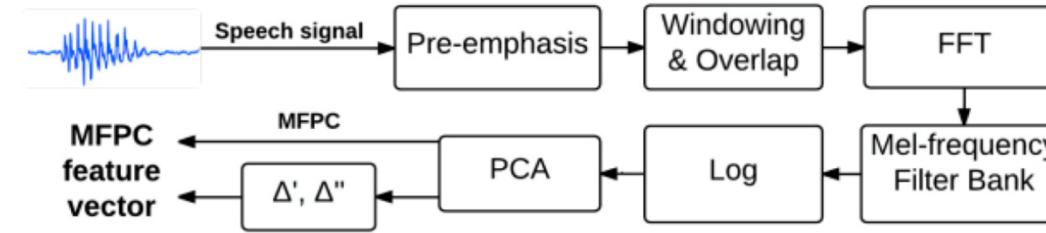
# System B (STC)

Feature extraction

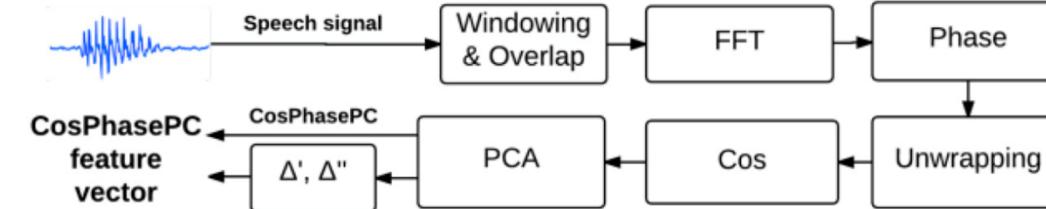
MFCC



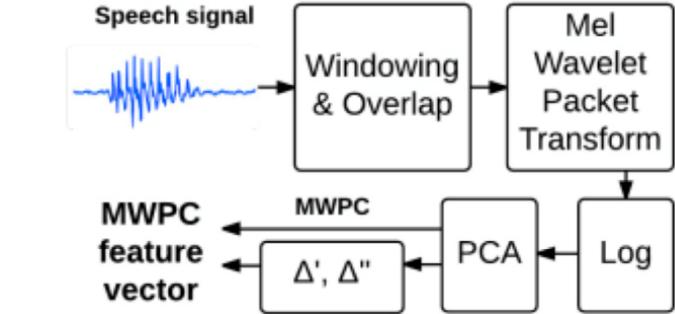
MFPC: Mel-Frequency Principle Coefficients



CosPhasePC: CosPhase Principle Coefficients



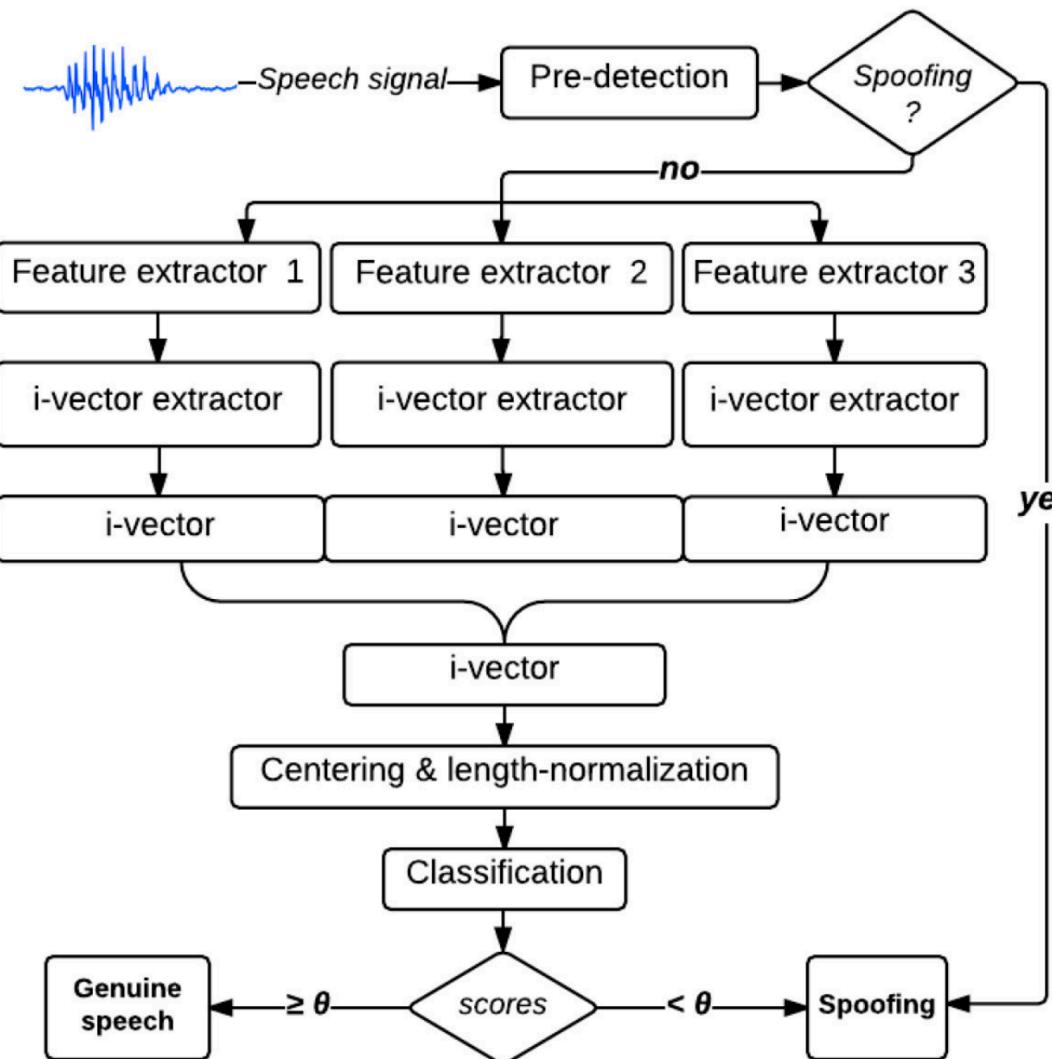
MWPC: Mel Wavelet Packet Coefficients



Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

# System B (STC)

Classifier

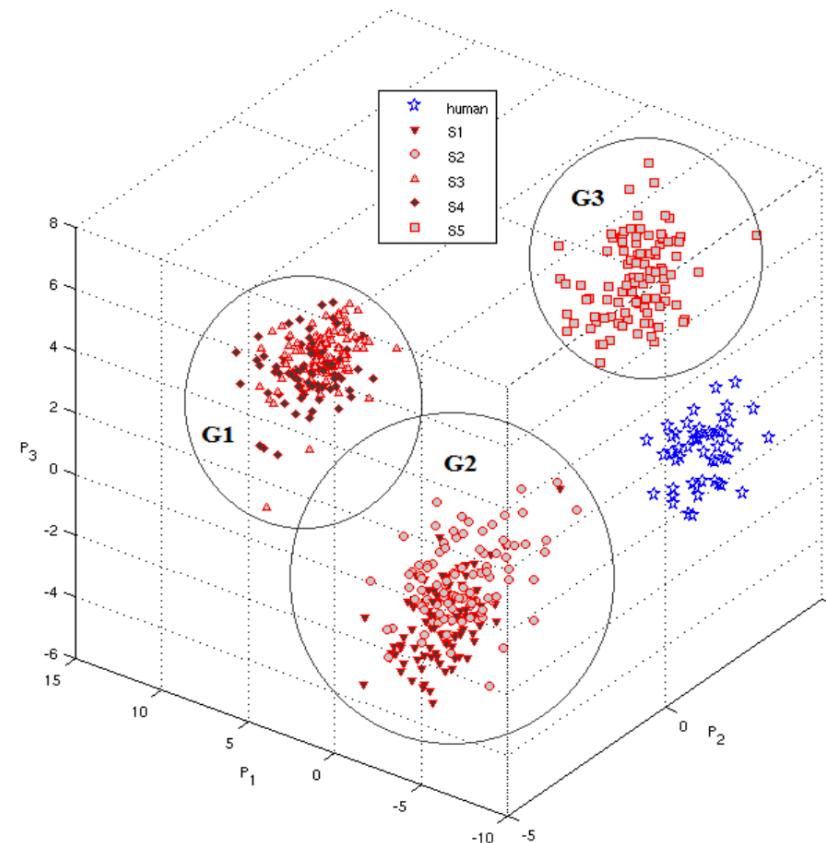


Sergey Novoselov, Alexandr Kozlov, Galina Lavrentyeva, Konstantin Simonchik, Vadim Shchemelinin, "STC Anti-spoofing Systems for the ASVspoof 2015 Challenge", arXiv:1507.08074, 2015

Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

# System B (STC): why success?

- MWPC: Mel Wavelet Packet Coefficients



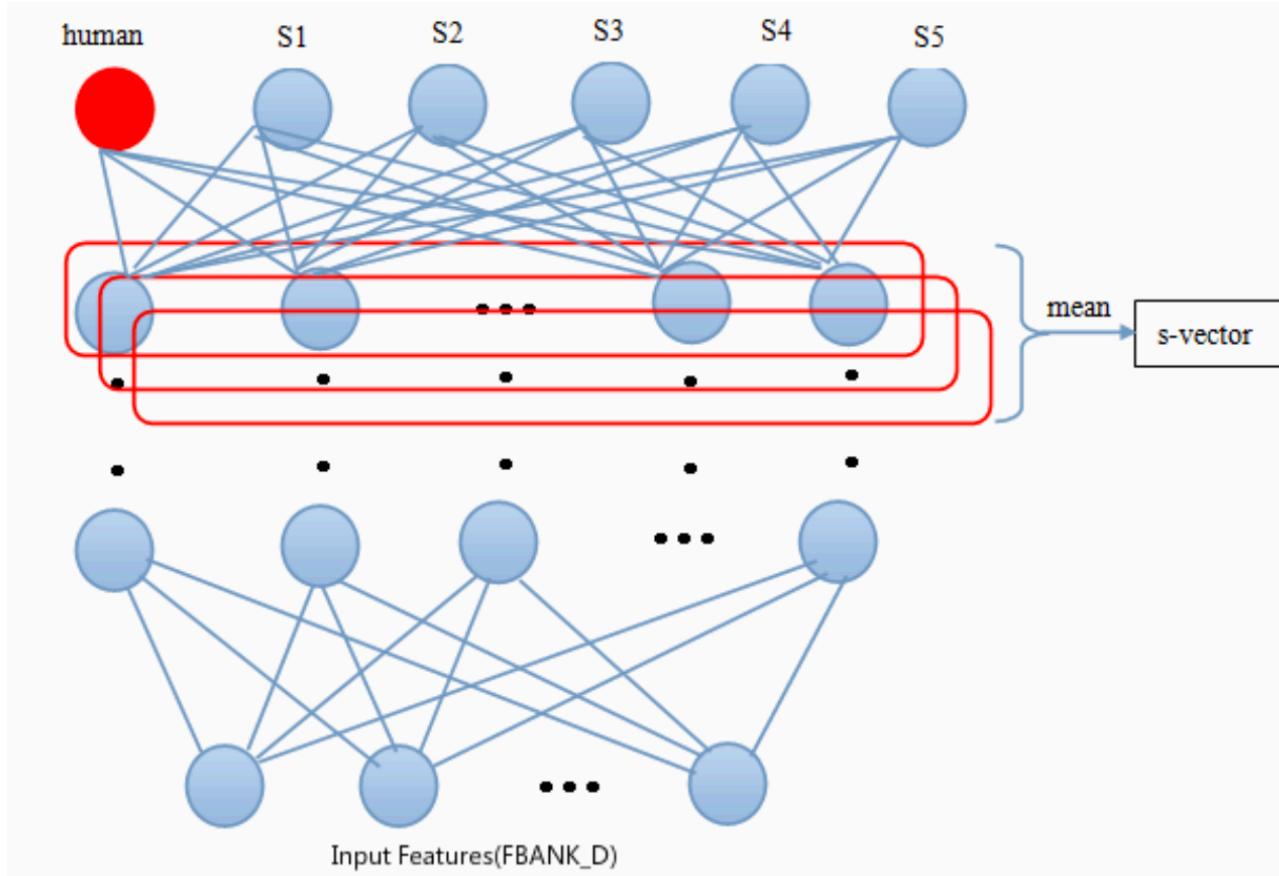
Sergey Novoselov, Alexandr Kozlov, Galina Lavrentyeva, Konstantin Simonchik, Vadim Shchemelinin, "STC Anti-spoofing Systems for the ASVspoof 2015 Challenge", arXiv:1507.08074, 2015

Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

# System C (SJTU)

Feature extraction

- A new feature: ‘s-vector’



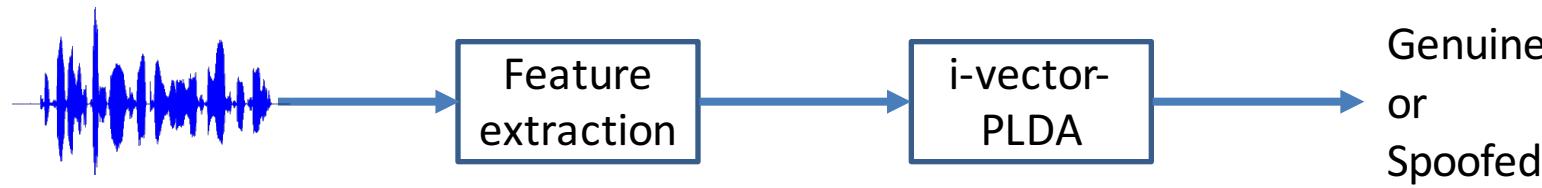
Nanxin Chen, Yanmin Qian, Heinrich Dinkel, Bo Chen, Kai Yu, "Robust Deep Feature for Spoofing Detection - The SJTU System for ASVspoof 2015 Challenge", Interspeech 2015

Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

# System C (SJTU)

Classifier

- i-vector-PLDA

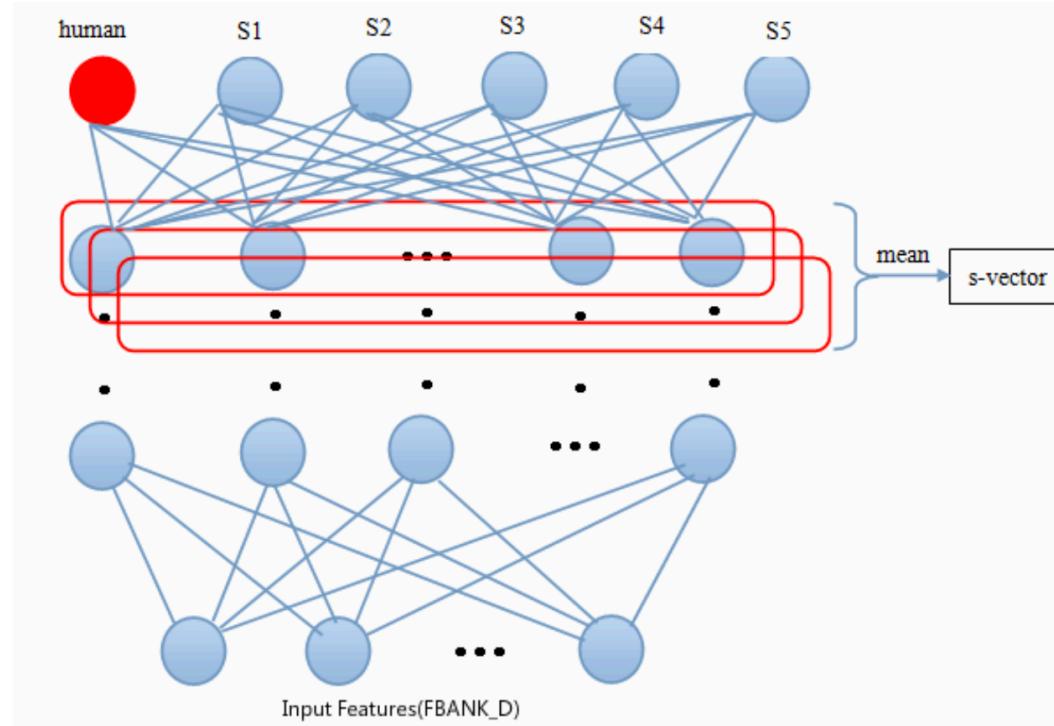


Nanxin Chen, Yanmin Qian, Heinrich Dinkel, Bo Chen, Kai Yu, "Robust Deep Feature for Spoofing Detection - The SJTU System for ASVspoof 2015 Challenge", Interspeech 2015

Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

# System C (SJTU): why success?

- Discriminative feature learnt by deep neural networks



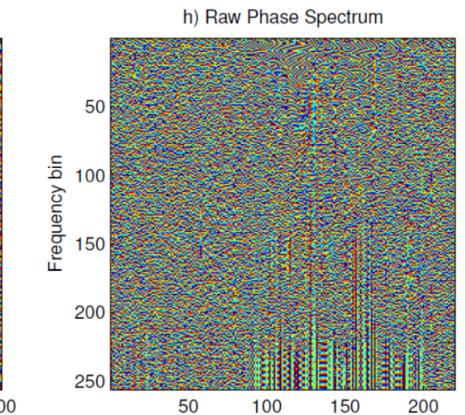
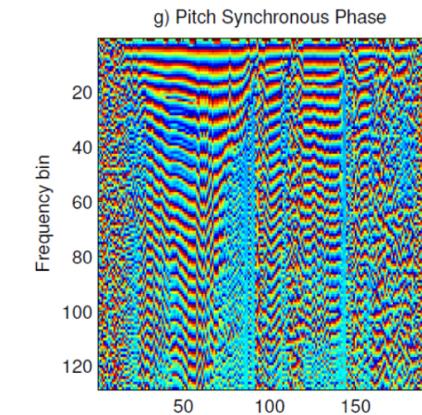
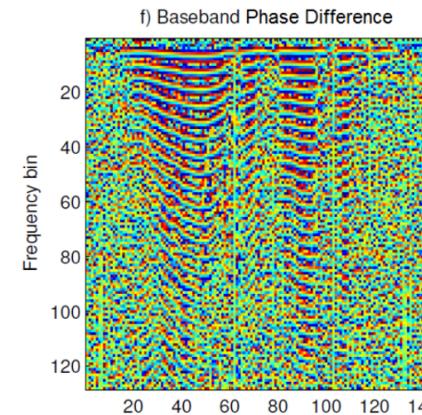
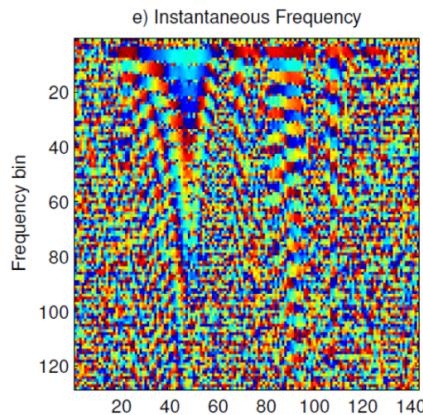
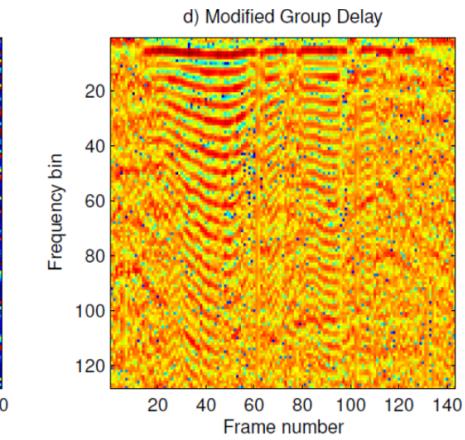
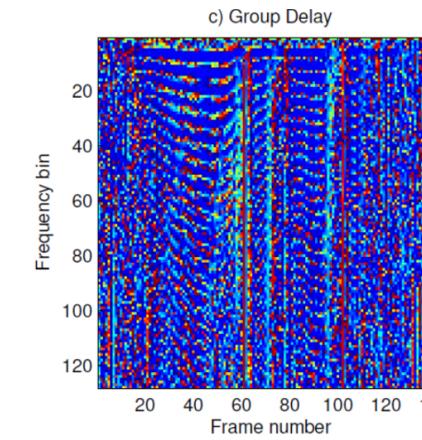
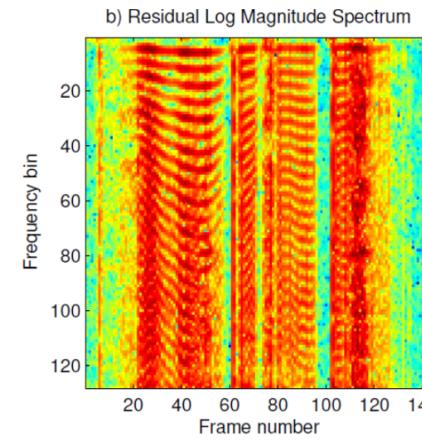
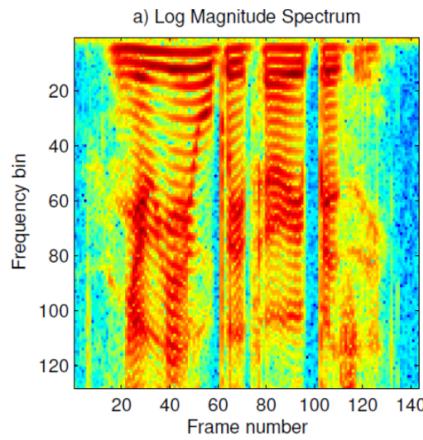
Nanxin Chen, Yanmin Qian, Heinrich Dinkel, Bo Chen, Kai Yu, "Robust Deep Feature for Spoofing Detection - The SJTU System for ASVspoof 2015 Challenge", Interspeech 2015

Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

# System D (NTU)

Feature extraction

- High-dimensional features: phase & magnitude



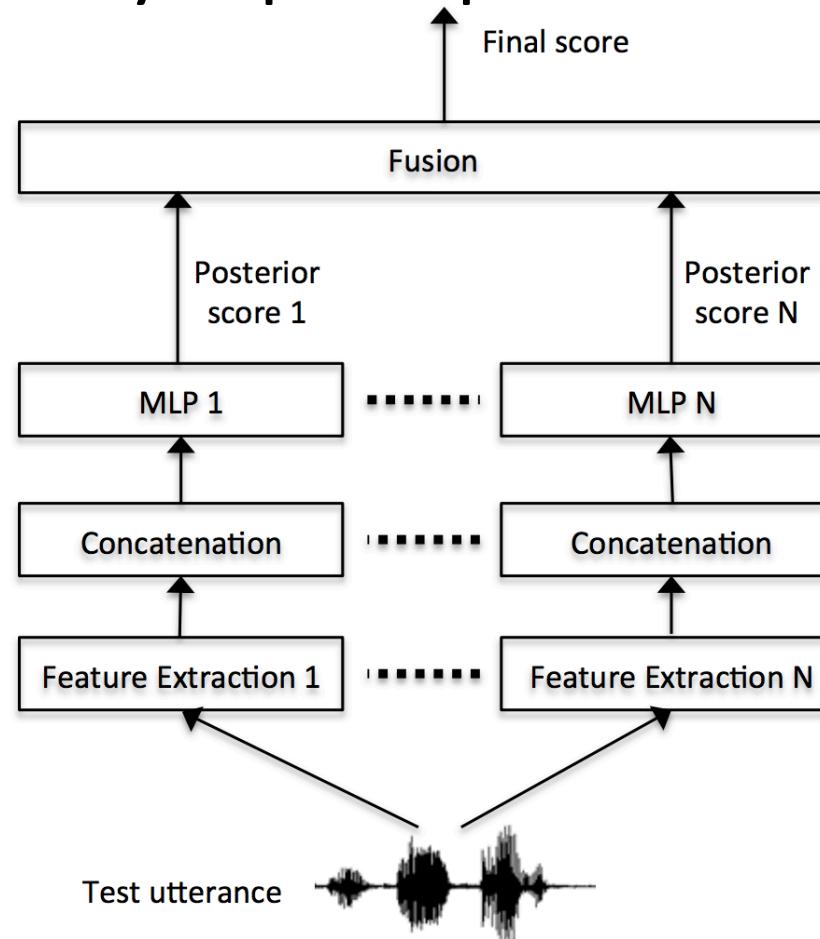
Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Eng Siong Chng, Haizhou Li, "Spoofing Speech Detection Using High Dimensional Magnitude and Phase Features: the NTU Approach for ASVspoof 2015 Challenge", Interspeech 2015

Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

# System D (NTU)

Classifier

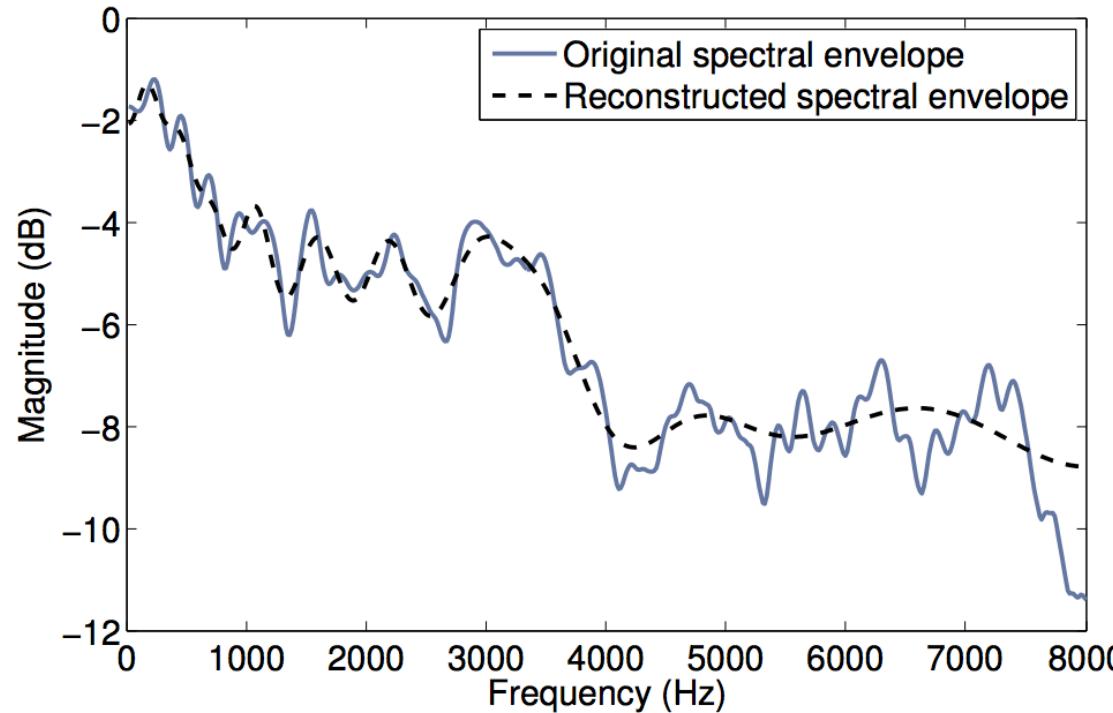
- MLP: multilayer perceptron with system fusion



Team	Average (all)
A	1.211
B	1.965
C	2.528
D	2.617

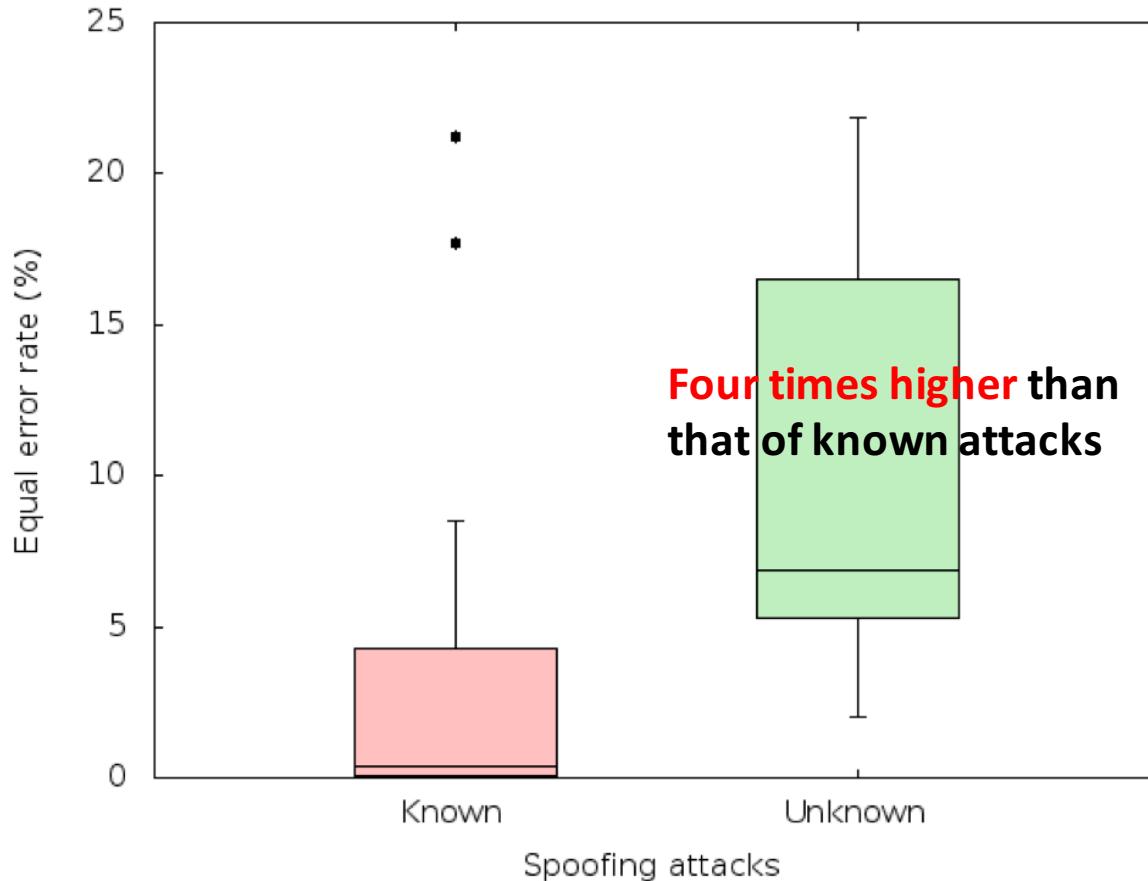
# System D (NTU): why success?

- High-resolution features that capture spectral & phase details, which are lost in VC and TTS

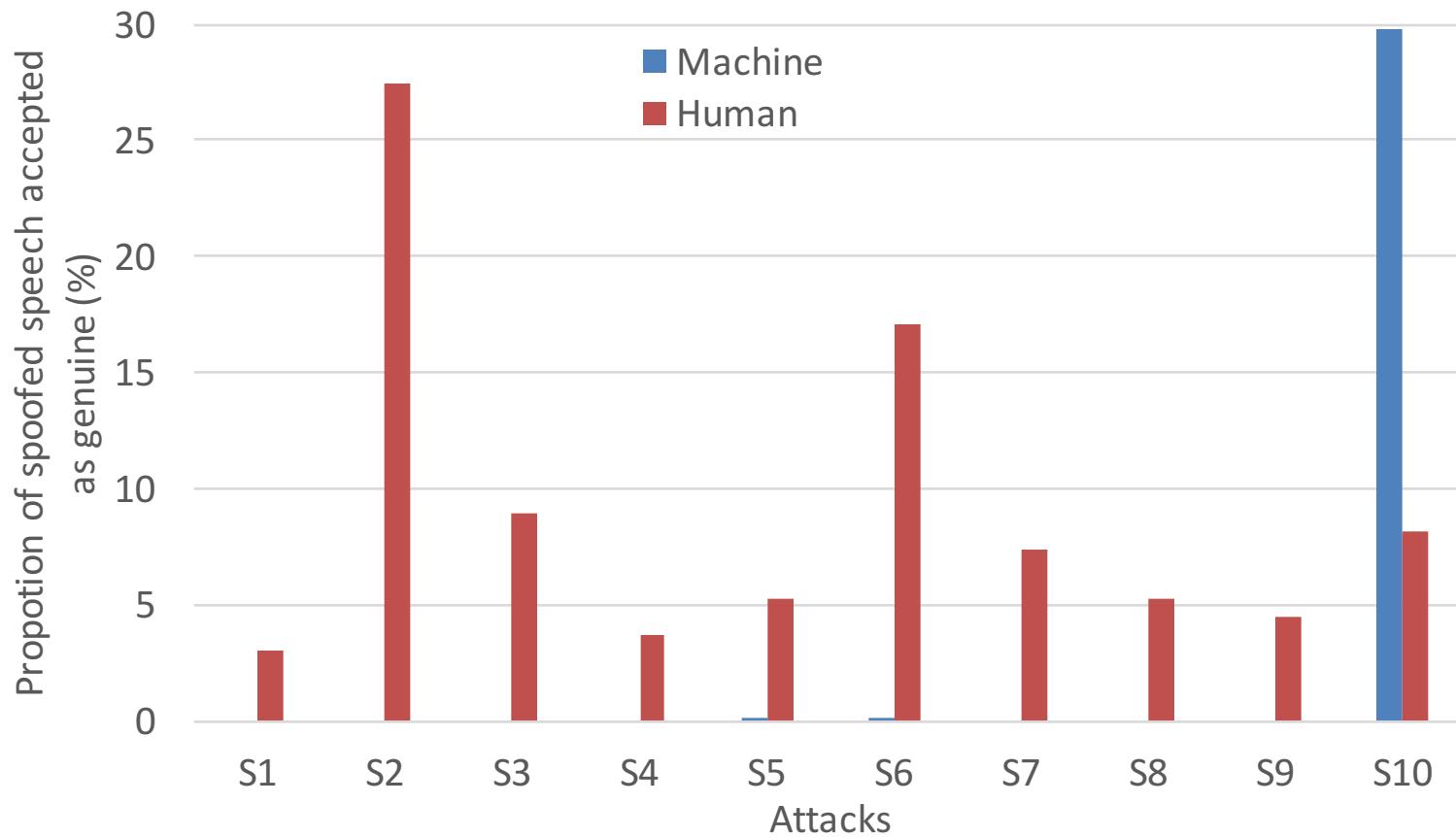


Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Eng Siong Chng, Haizhou Li, "Spoofing Speech Detection Using High Dimensional Magnitude and Phase Features: the NTU Approach for ASVspoof 2015 Challenge", Interspeech 2015

# Good news vs bad news



# Human vs Machine



S2: VC – slope shifting

S6: VC – GMM-based conversion with global variance enhancement

S10: TTS – unit selection

# Special session @ INTERSPEECH 2015



Zhizheng Wu, Tomi Kinnunen, Nicholas Evans and Junichi Yamagishi, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge", IEEE Signal Processing Society Speech and Language Technical Committee Newsletter (SLTC Newsletter), 20 November 2015

# Play with ASVspoof database?

- ASVspoof database
  - <http://dx.doi.org/10.7488/ds/298>
- Evaluation plan:
  - <http://www.spoofingchallenge.org/asvSpoof.pdf>
- INTERSPEECH summary paper
  - [http://www.spoofingchallenge.org/is2015\\_asvspoof.pdf](http://www.spoofingchallenge.org/is2015_asvspoof.pdf)

# Summary

- The first challenge is highly successful in attracting significant participation
  - At least 10 companies are interested in the database (post-challenge)
- Most of the participants achieved good results on known attacks, however, many of them got higher error rates on unknown attacks
- There is still a long way to go towards a real generalised countermeasure

# Outline

## Part 1

1. Introduction
2. Speaker verification
3. Speech synthesis
4. Voice conversion
5. Q&A

## Part 2

6. Spoofing
7. Countermeasures
8. ASVspoof 2015
9. Future
10. Q&A



# Future directions

- Generalised countermeasures
- Text-dependent verification
- Replay attacks
- Different vocoders
- Noise and channel variability
- Combination with ASV
- Speaker dependent countermeasures
- Combined spoofing attacks and fused countermeasures
- Evaluation metrics
- ASVspoof 2017

# Future information

- Challenge website:
  - <http://www.spoofingchallenge.org/>
  - System descriptions are available at the website
- SLTC newsletter: Nov. 2015
  - <http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2015-11/2015-11-ASVspoof/>

## Challenge papers and results

Zhizheng Wu, Tomi Kinnunen, Nicolas Evans, Junichi Yamagishi, Cemal Hanilci, Md Sahidullah, Aleksandr Sizov, "ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge", Interspeech 2015 [\[PDF\]](#)

Md Jahangir Alam, Patrick Kenny, Gautam Bhattacharya, Themos Stafylakis, "Development of CRIM System for the Automatic Speaker Verification Spoofing and Countermeasures Challenge 2015", Interspeech 2015 [\[PDF\]](#)

Nanxin Chen, Yanmin Qian, Heinrich Dinkel, Bo Chen, Kai Yu, "Robust Deep Feature for Spoofing Detection - The SJTU System for ASVspoof 2015 Challenge", Interspeech 2015 [\[PDF\]](#)

Artur Janicki, "Spoofing Countermeasure Based on Analysis of Linear Prediction Error", Interspeech 2015 [\[PDF\]](#)

Yi Liu, Yao Tian, Liang He, Jia Liu, Michael T. Johnson, "Simultaneous Utilization of Spectral Magnitude and Phase Information to Extract Supervectors for Speaker Verification Anti-spoofing", Interspeech 2015 [\[PDF\]](#)

Sergey Novoselov, Alexandr Kozlov, Galina Lavrentyeva, Konstantin Simonchik, Vadim Shchemelinin, "STC Anti-spoofing Systems for the ASVspoof 2015 Challenge", arXiv:1507.08074, 2015 [\[PDF\]](#)

Tanvina B. Patel, Hemant A. Patil, "Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural vs. Spoofed Speech", Interspeech 2015 [\[PDF\]](#)

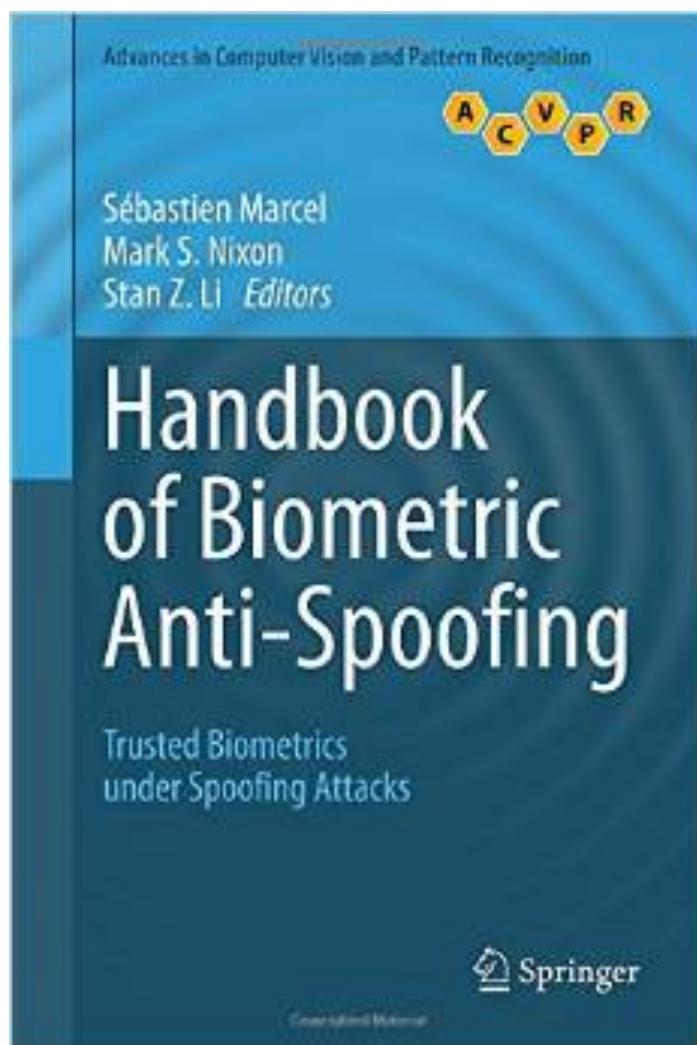
Jon Sanchez, Ibon Saratxaga, Inma Hernaez, Eva Navas, Daniel Erro, "The AHOLAB RPS SSD Spoofing Challenge 2015 submission", Interspeech 2015 [\[PDF\]](#)

Jesus Villalba, Antonio Miguel, Alfonso Ortega, Eduardo Lleida, "Spoofing Detection with DNN and One-class SVM for the ASVspoof 2015 Challenge", Interspeech 2015 [\[PDF\]](#)

Longbiao Wang , Yohei Yoshida, Yuta Kawakami, Seiichi Nakagawa, "Relative phase information for detecting human speech and spoofed speech", Interspeech 2015 [\[PDF\]](#)

Shitao Weng, Shushan Chen, Lei Yu, Xuewei Wu, Weicheng Cai, Zhi Liu, Ming Li, "The SYSU System for the Interspeech 2015 Automatic Speaker Verification Spoofing and Countermeasures Challenge", arXiv:1507.06711, 2015 [\[PDF\]](#)

Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Eng Siong Chng, Haizhou Li, "Spoofing Speech Detection Using High Dimensional Magnitude and Phase Features: the NTU Approach for ASVspoof 2015 Challenge", Interspeech 2015 [\[PDF\]](#)



Speech Communication

Volume 66, February 2015, Pages 130–153



## Spoofing and countermeasures for speaker verification: A survey

Zhizheng Wu<sup>a</sup>, Nicholas Evans<sup>b</sup>, Tomi Kinnunen<sup>c</sup>, Junichi Yamagishi<sup>d, e</sup>, Federico Alegre<sup>b</sup>, Haizhou Li<sup>a, f</sup>

[Show more](#)

Choose an option to locate/access this article:



[Get Full Text Elsewhere](#)

[doi:10.1016/j.specom.2014.10.005](#)

[Get rights and content](#)

### Abstract

While biometric authentication has advanced significantly in recent years, evidence shows the technology can be susceptible to malicious spoofing attacks. The research community has responded with dedicated countermeasures which aim to detect and deflect such attacks. Even if the literature shows that they can be effective, the problem is far from being solved; biometric systems remain vulnerable to spoofing. Despite a growing momentum to develop spoofing countermeasures for automatic speaker verification, now that the technology has matured sufficiently to support mass deployment in an array of diverse applications, greater effort will be needed in the future to ensure adequate protection against spoofing. This article provides a survey of past work and identifies priority research directions for the future. We summarise previous

# Acknowledgements

- Thank the following colleagues from providing spoofing materials
  - Dr. Daisuke Saito from University of Tokyo, Japan
  - Prof. Tomoki Toda from Nagoya University, Japan
  - Prof. Zhen-Hua Ling from University of Science and Technology of China
  - Mr Ali Khodabakhsh and Dr. Cenk Demiroglu from Ozyegin University, Turkey
- Protocol validation (conduct pilot evaluation)
  - Dr. Md Sahidullah, Dr. Cemal Hanilci, Mr. Aleksandr Sizov from University of Eastern Finland

# Acknowledgements

- The work was partially supported by
  - EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology)
  - Academy of Finland (project no. 253120 and 283256)
  - OCTAVE project which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 647850.

# Outline

## Part 1

1. Introduction
2. Speaker verification
3. Speech synthesis
4. Voice conversion
5. Q&A

## Part 2

6. Spoofing
7. Countermeasures
8. ASVspoof 2015
9. Future
10. Q&A



