David van Leeuwen

TNO SRE-2005 submission

Defence, Security and Safety

TNO | Knowledge for business

The bottom line:
improvements from last year's system

Development test 'half 2004':	min	act	EER(%)
• TNO 2004 GMM system	0.533	0.551	14.5
• + 2004 tnorm models	–0.021		–1.28
• + feature mapping	–0.021		–1.59
• + feature mapping + 2004 tnorm	–0.072		–2.95
• 512 → 2048 Gaussians	–0.078		–2.50
• TNO 2004 SVM system	0.574	0.643	14.3
• + 2004 tnorm	–0.082		–1.97
• + feamap-2048, 2004 tnorm	–0.097		–2.48

The bottom line:
improvements from last year's system

System Fusing, all trials, 1side-1side

Dev 'half 2004'	min	act	EER(%)
• TNO 2004 5 subsystems linear fuse	0.517	0.574	13.9
• TNO 2005 gmm+svm linear fuse	–0.067	–0.119	–2.53
• lnknet fuse	–0.066	–0.120	–2.52
• + word *n*gram	–0.077	–0.126	–2.35
• + 3×SDV system	–0.144	–0.198	–3.98
• + PAV	–0.145	–0.198	–3.92
• which makes	0.372	0.376	9.89
• SRE 2005	0.271	0.282	7.90

Development test data

- NIST SRE 2004
- split in two halves per sex
  - ‘train’ half
    - T-norm models, calibration, PAV, …
  - ‘development test’ half
    - optimization of parameters

  - Random samples, but under constraints
    - difference sample min DCF < 0.01
    - difference sample EER < 0.5 %
    - for both 2004 GMM(1024) and SVM system
      - ~ 75 attempts of split
      - σ(EER) ~ 0.8%, σ(mDCF) ~ 0.025

Feature mapping

- After Doug Reynolds, ICASSP 2003
- 8 channels for ‘root UBM’
- 2 sex $\times$ 4 microphone/channel types
  - Switchboard 2 phase 2 landline, MIT-LL classification
    - carbon/button
    - electret
  - NIST SRE 2001–2003
    - GSM
    - CDMA
- 80 speakers/channel, but
  - only 50+61 carbon/button m+f
- 591 speakers in total
- MAP adaptation of means from root UBM to each channel

Features and models

- Frame energy based E_{max} – 30 dB speech detection
  - yield ~ 30%
- 12 PLP coefficients + energy + delta/7, normalized,
  - no more feature warping
- Feature mapping using 8 channels, normalized
- UBM/GMM using 2048 mixtures
  - root UBM as UBM
- SVM using feature mapped inputs
  - 591 feature mapped background speakers

T-norm models
or how to be inefficient

- 315 speaker models from ‘train’ half 2004
  - 155 different speakers
  - possibly 2 different channels per speaker?

- T-normalization sex-independent
  - applied using all 315 models

- Various samples of this set tried, but failed in performance
  - cohort selection using distance measure
  - top or bottom *N* models per trial
  - sex-dependent
  - accent-dependent

Word *n*-gram

- Background words from ‘train’ half 2004 ASR output
- Vocabulary restricted to words with frequency > 9
  - limits size LM
- Background bigram LM,
  - constant discount 0.1
  - SRILM toolkit

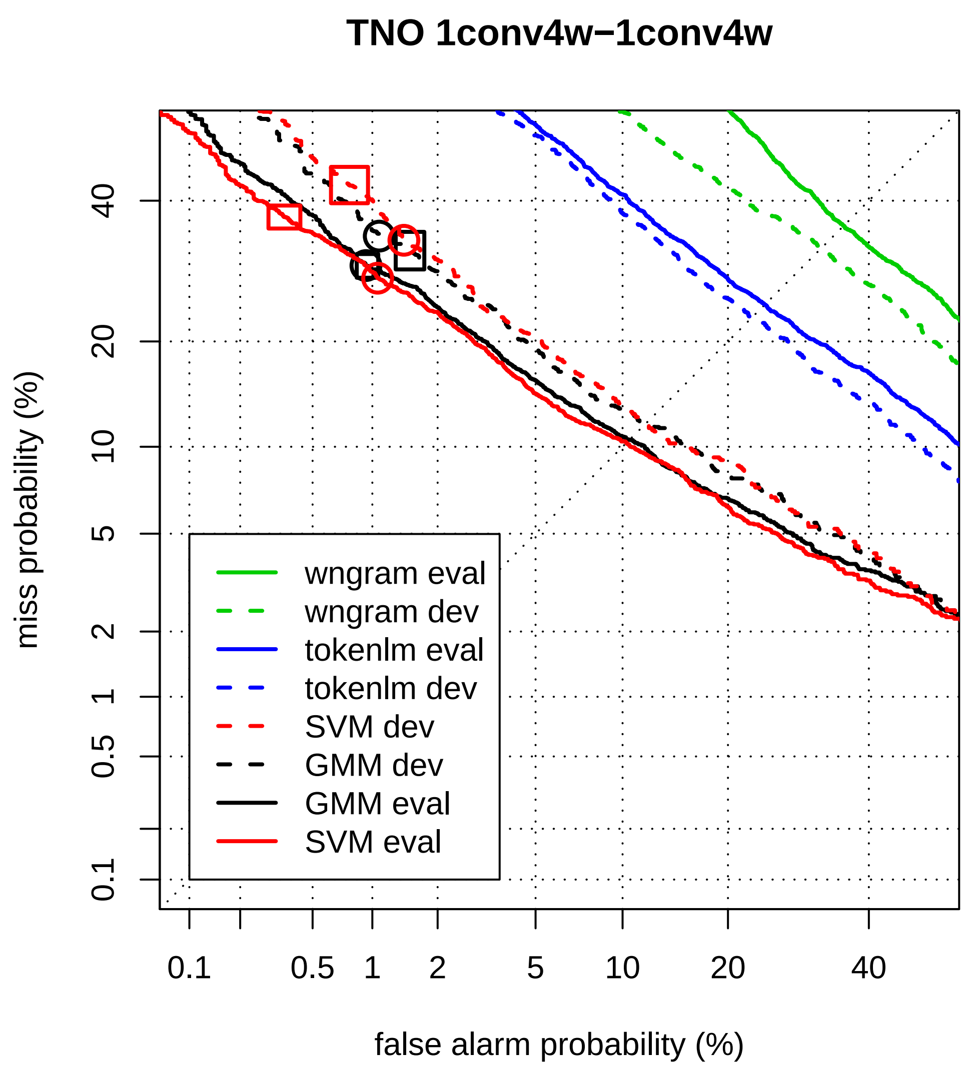
- Train:
  - build LM on ASR words,
  - mix (interpolate) LM with background LM for smoothing

- Test log likelihood score measure:
  - minus perplexity of test ASR word string on speaker LM

TokenIm or ID-*n*-gram

- Basic token per frame:
  - index of most likely Gaussian
    - 512 Gaussians used
  - ‘free’ side-information in feature mapping process

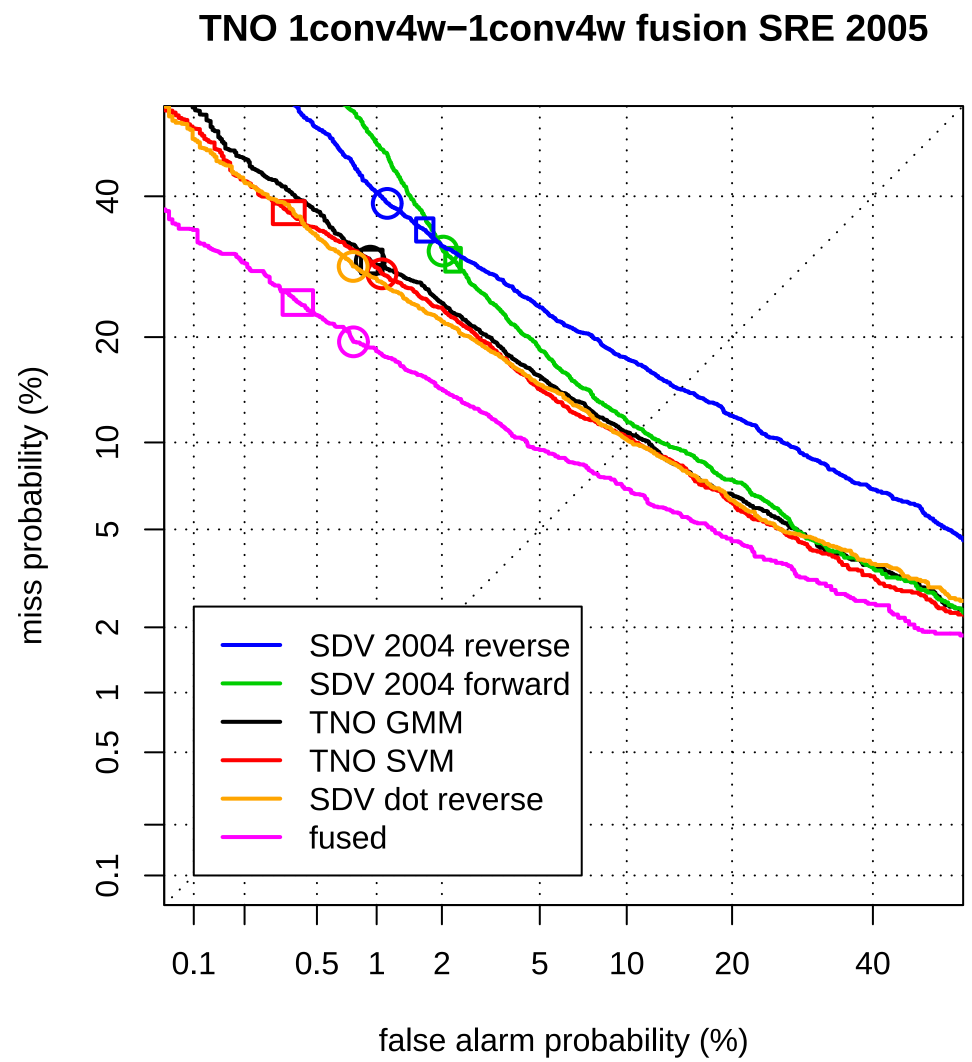
- Otherwise identical to word *n*-gram


TNO 1conv4w−1conv4w
miss probability (%)
false alarm probability (%)
0.1
0.5
1
2
5
10
20
40
wngram eval
wngram dev
tokenlm eval
tokenlm dev
SVM dev
GMM dev
GMM eval
SVM eval

Fusing

- This year, investigated `lnknet` software rather than linear fusion
- `lnknet`
  - needs training/calibration data set
    - we (re-)used ‘train’ half $\Rightarrow$ questionable
  - accepts *prior class probability*
    - here: ‘effective prior odds’ 9.9:1
  - makes decision on ‘posterior odds > 1’
    - we used PAV score$\rightarrow$likelihood ratio mapping

- Biggest gain in performance from fusing with 3 SDV systems
  - SDV eigenchannel forward (SDV3)
  - SDV eigenchannel reverse (SDV4)
  - SDV adapted supervector dot product reverse (SDV6)


TNO 1conv4w−1conv4w fusion SRE 2005
miss probability (%)
0.1
0.5
1
2
5
10
20
40
false alarm probability (%)
0.1
0.5
1
2
5
10
20
40
SDV 2004 reverse
SDV 2004 forward
TNO GMM
TNO SVM
SDV dot reverse
fused

Pool Adjacent Violators algorithm (PAV)
from the creative brain of Niko Brümmer

- converts scores to likelihood ratios
  - uses training scores and truths
    - here ‘train’ half 2004
  - calibrates likelihood ratios to application type
    - here ‘effective prior odds’ 1/9.9

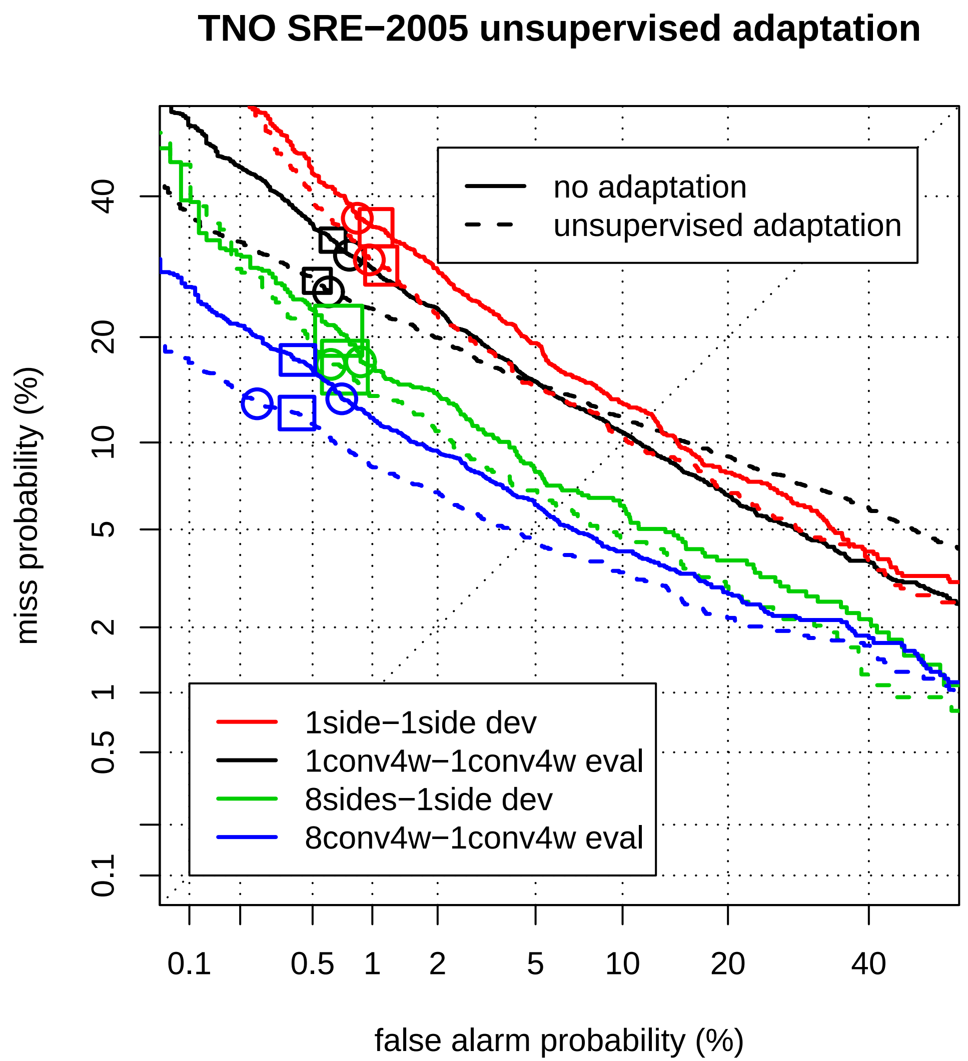
$$\mathcal{O}_{\text{eff}} = \frac{C_{\text{miss}}}{C_{\text{FA}}} \frac{P_{\text{target}}}{1 - P_{\text{target}}}$$

    - set decision threshold to log(9.9)~2.29
  - but output is also calibrated for other application types

Unsupervised adaptation in 2005:
quite a hassle but it seems to work now

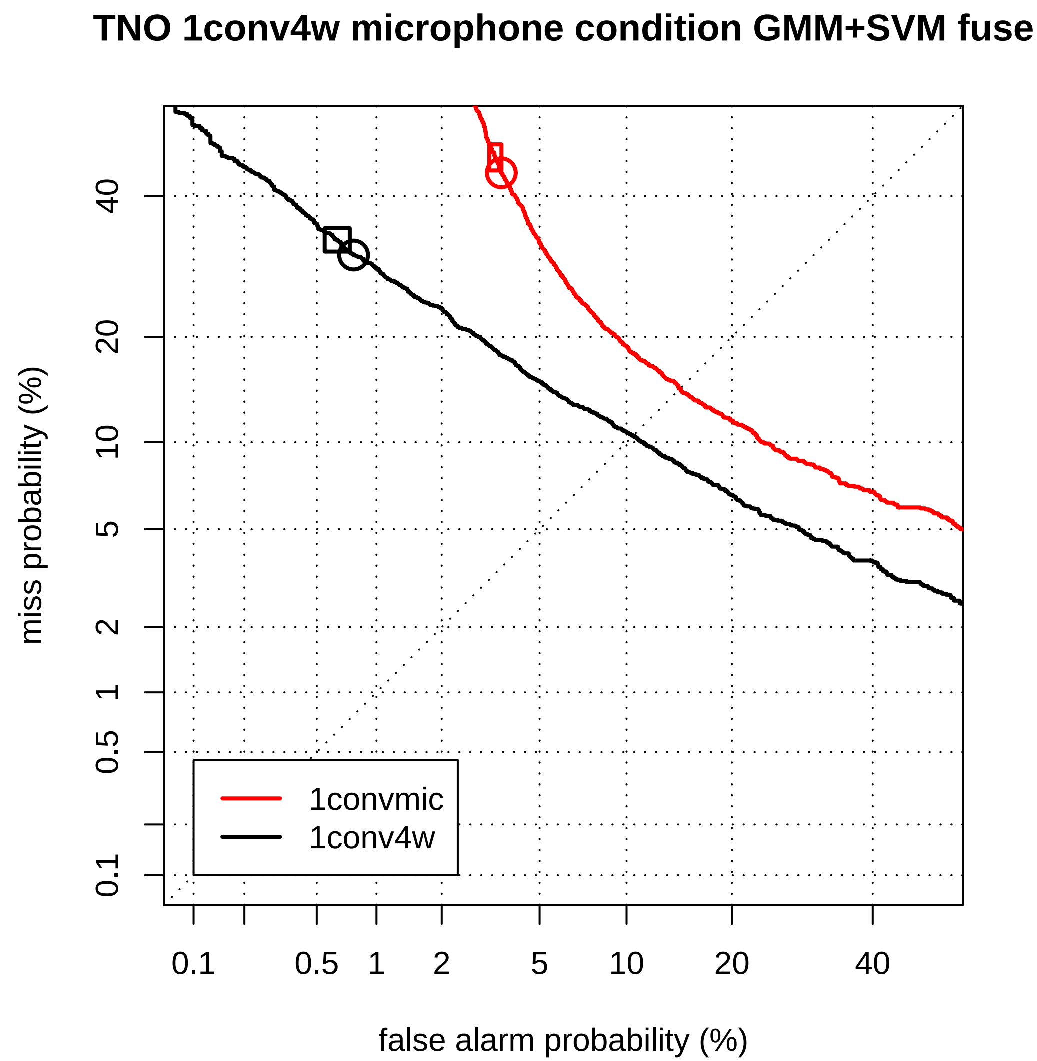
- GMM: same adaptation principle as in 2004 (Claude Barras, Odyssey 2004)
  - process trials sequentially
  - adapt speaker model with test segment if T-normed score > *a*, using relevance *r*
- SVM: new this year for 8conv4w-1conv4w
  - add test segment to positive examples and retrain SVM, if T-normed score > *a*

train condition	1conv4w		8conv4w	
system	*a*	*r*	*a*	*r*
GMM-512	3.5	24	-	-
GMM-2048	5	8	4	16
SVM	-		4	


TNO SRE−2005 unsupervised adaptation
no adaptation
unsupervised adaptation
1side−1side dev
1conv4w−1conv4w eval
8sides−1side dev
8conv4w−1conv4w eval
miss probability (%)
false alarm probability (%)
0.1
0.5
1
2
5
10
20
40

1convmic test condition

- Feature mapping approach
- used available training material for microphone data distributed by NIST
- use 20 dB dynamic range for speech/non-speech detector
  - rather than 30 dB for telephone signals
- 8 channels, as from training
- Train microphone channel models from root UBM
- In testing, classify each test segment as one of microphone channels, map features back to ‘telephone feature space’
- Otherwise the same as in normal speaker detection
- No development test material available
  - T-normalization and fixed threshold of 3.0
- ASR output not used


TNO 1conv4w microphone condition GMM+SVM fuse
miss probability (%)
false alarm probability (%)
0.1
0.5
1
2
5
10
20
40
1convmic
1conv4w

Other submissions

- This year, devtest material was available for all other conditions
  - 8conv4w-1conv4w (6/13)
    - influence of more training, word *n*-gram, SVM adaptation
  - 8conv4w-10sec4w (1/6)
    - robustness against short test segments
  - 10sec4w-10sec4w (1/10)
    - robustness against short training segments
- Not investigated, but possibly interesting
  - 10sec4w-1conv4w
    - unsupervised adaptation mode
    - no adaptation: use ‘reverse trick’ (SDV) or symmetric measure (e.g., SVM)

Conclusions

- Teaming up with alternative system developer is useful
- Feature mapping is good idea
  - microphone test conditions
  - SVM input features
- Difficult to get good results with cohort T-norming
  - best results with full set, but slow
  - for some systems, T-norming not important in fusing
- Lnknet + PAV can work together for fusion and calibration
  - PAV more consistently better in calibration (not for 1c-1c…)
- Speaker adaptation in NIST SRE tricky
  - evaluation priors not realistic to verification application
  - optimal parameter settings not very robust against SRE year
- SRE 2005 ‘easier’ than SRE 2004