Brno University of Technology system for NIST SRE 2008

Speech@FIT

Lukáš Burget, Michal Fapšo, Valiantsina Hubeika, Ondřej Glembek, Martin Karafiát, Marcel Kockmann, Pavel Matějka, Petr Schwarz and Honza Černocký

http://speech.fit.vutbr.cz, speech@fit.vutbr.cz

Submitted systems

BUT01 - primary (3 systems) - Channel and language side information in fusion
- FA-MFCC13→39
- FA-MFCC20→60
- SVM-MLLR

BUT02 - (3 systems) - The same as BUT01 but no side information in fusion

BUT03 - (2 systems) - Channel and language side information in fusion
- FA-MFCC13→39
- FA-MFCC20→60

FA-MFCC13→39 system

- MAP adapted UBM with 2048 Gaussian components - Single UBM trained on Switchboard and NIST 2004,5 data
- Features: 12 MFCC + C0 (20ms window, 10ms shift)
- Short time Gaussianization - Rank of the current frame coefficient in 3sec window transformed by inverse Gaussian cumulative distribution function.
- Delta + double delta + triple delta coefficients - Together 52 coefficients, 12 frames context
- HLDA (dimensionality reduction from 52 to 39)
- Factor Analysis Model – gender independent
  - 300 eigenvoices (Switchboards, NIST 2004,5)
  - 100 eigenchannels for telephone speech (NIST 2004,5 tel. data)
  - 100 eigenchannels for microphone speech (NIST 2005 mic. data)
- ZT-norm – gender dependent

FA-MFCC20→60 system

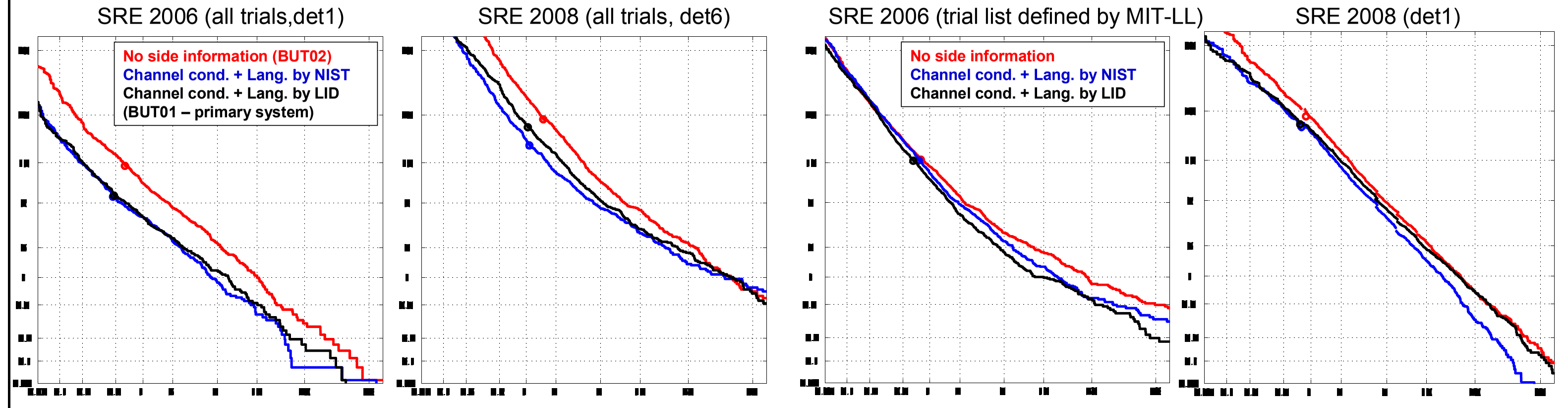
- The same as FA-MFCC13→39 with the following differences:
  - 60 dimensional features are: 19 MFCC + Energy + deltas + double deltas (no HLDA)
  - Two gender dependent Factor Analysis models

SVM – MLLR system

- Linear kernels, Rank normalization, LibSVM C++ library [Chang2001], Pre-computed Gram matrices
- Features are MLLR transformations adapting LVCSR system (developed within AMI project) to speaker of given speech segment
- Estimation of MLLR transformations makes use of the ASR transcripts provided by NISTCascade of CMLLR and MLLR
  - 2 CMLLR transformation (silence and speech)
  - 3 MLLR transformation (silence and 2 phoneme clusters)
- Silence transformations are discarded for SRE
- Supervector = 1 CMLLR + 2 MLLR = 3*392+3*39=4680
- Impostors: NIST 2004 + mic data from NIST 2005
- ZT-norm: speakers from NIST 2004

Side info based calibration and fusion

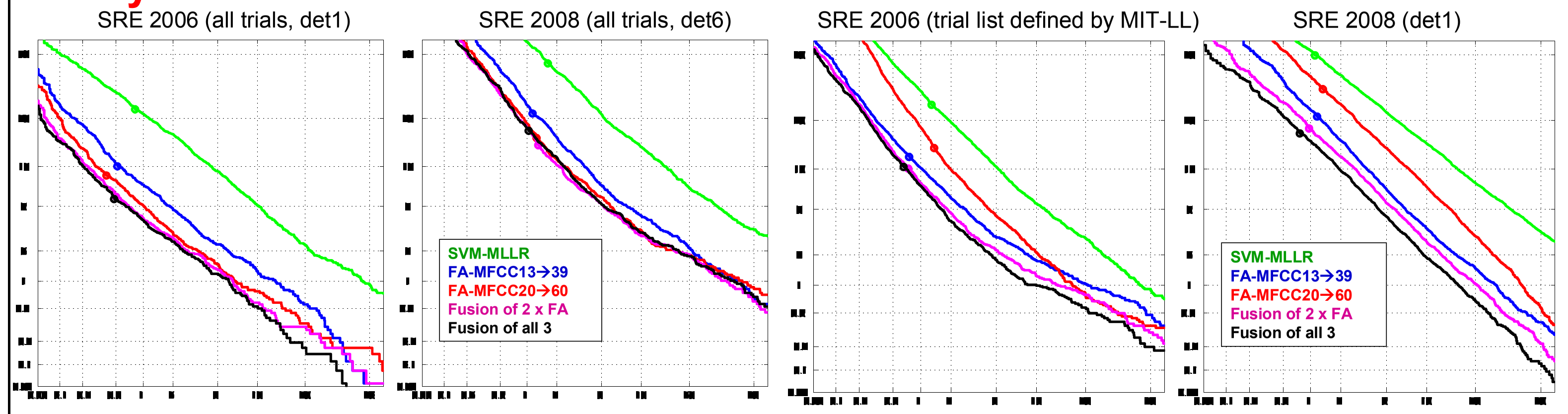
- **For each system:**
  1. Split trials by channel condition and calibrate scores using linear logistic regression (LLR) in each split separately
  2. Split trials according to English/non-English decision and calibrate scores using LLR in each split separately
- **Fuse the calibrated scores of all subsystems using LLR without making use of any side information**
- Side information for each trial is given by its hard assignment to classes:
  - Trial channel condition provided by NIST: tel-tel, tel-mic, mic-tel, mic-mic
  - English/non-English decision given by our LID system
- FoCal Bilinear toolkit by Niko Brummer was used, no use of its extensions over standard LLR.



Tel-tel trials: Use of side information is helpful and some improvement can be obtained by relying on language information provided by NIST instead of more realistic LID system

Mic-mic trials: Use of side information allowed to the same unchanged subsystems for all the channels

Subsystems and fusion



Tel-tel trials: single FA-MFCC20→60 performs almost as well as the fusion

Mic-mic trials: FA-MFCC20→60 fails to perform well, fusion is beneficial. FA-MFCC13→39 system outperforms FA-MFCC20→60 system having 3x more parameters, which is possibly too over-trained to telephone data.

Other systems that did not make it to our NIST submission…

- GMM with eigenchannel adaptation
- SVM-GMM 2048 + NAP
- SVM-GMM 2048 + ISV (Inter Session Variability modeling)
- SVM-GMM 2048 + ISV derivative (Fisher) kernel
- SVM-GMM 2048 + IVS based on FA-MFCC13→39
- FA modeling prosodic and cepstral contours
- SVM on phonotactics – counts from Binary decision trees
- SVM on soft bigram statistics collected on cumulated posteriograms (matrix of posterior probability of phonemes for each frame)

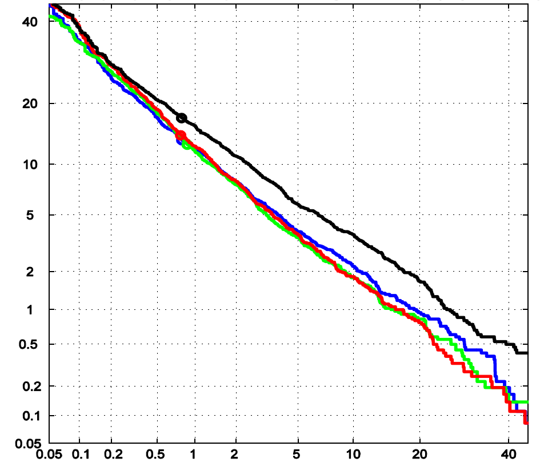
Not used in the submission to NIST as they did not bring complementary information

SVM –GMM subsystems

MFCC 12 + C0 + delta + doubledelta + tripple delta
- HLDA (dimensionality reduction 52→39)
- T-norm with 2004 data
- Impostor speakers – NIST 2004 data + microphone data from NIST 2005

Different flavours:
- UBM 512G + NAP
- UBM 2048G + NAP (similar results to 512G)
- UBM 2048G + ISV
- 2048G + ISV – derivative kernels
- UBM 2048G + ISV + eigenvoices

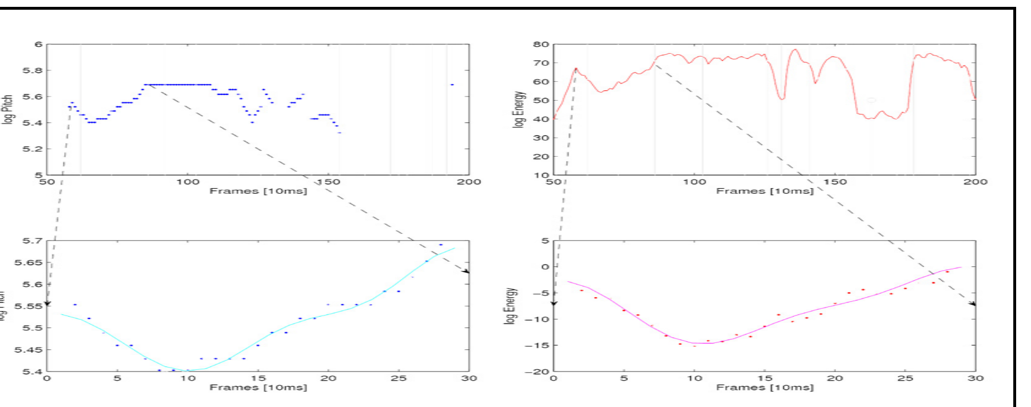
NAP and eigen channel compensation based on 50 vectors derived from NIST 2004 data and 50 vectors from NIST 2005 microphone data



System/SRE06 det3 (Eng. only)	EER [%]
512G + NAP	2.46
2048G + ISV	2.21
2048G + ISV derivative kernels	2.17
2048G + ISV + eigen-voices	3.69

Prosodic subsystems

- 2 systems modelling temporal contours based on pseudo-syllables
- Syllable segmentation based on phoneme recognizer and pitch
- Basic short-time features: Pitch, Energy and MFCCs



- Syllable-contours are represented by 4 DCT coefficients for each basic feature type
- Only voiced segments (based on pitch) are used, min. 40ms
- System 1:
  Duration and Contours for Pitch and Energy (13 dim. features)
- System 2:
  Duration and Contours for Pitch, Energy and 12 MFCCs (57 dim. features)

GMM-Eigenchannel vs. Factor Analysis Framework (gender independent):

	no CC	Model Domain	Feature Domain	Factor Analysis
System 2 512 Gaussians	9,44	9,06	9,06	5,91

Final systems:
- Gender dependent, zt-norm
- System 1: 50 eigenvoices and 40 eigenchannel (20 tel/20 mic) per gender
- System 2: 150 eigenvoices and 70 eigenchannel (35 tel/35 mic) per gender

	EER [%]
System 1 512 Gaussians	14,52
System 2 512 Gaussians	3,74

All results on SRE2006 1conv4w-1conv4w English only

Phonotactic sub-systems

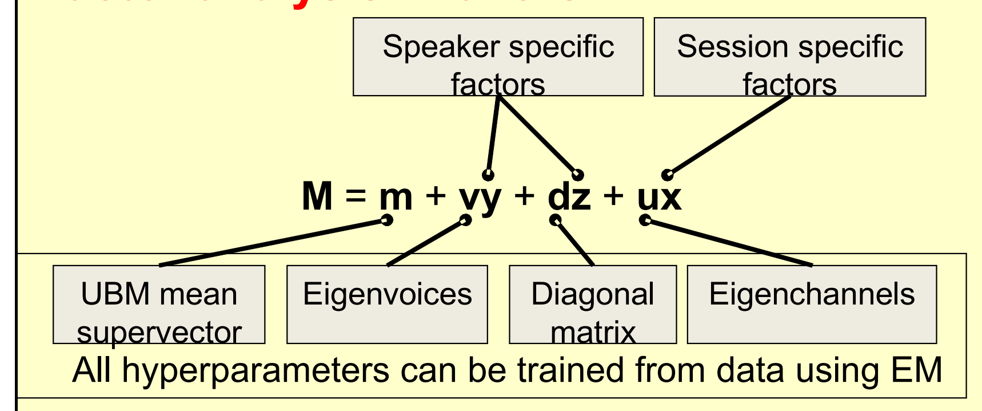
Phoneme posterior based system:
- Phoneme state posteriors for each frame
- Averaged phoneme state posteriors based on 1-best segmentation
- 3 states of each phoneme summed to one posterior (62 phonemes)
- Phoneme bigram statistics for the whole utterance
- Normalized super-vector for SVM (3844 features/vector)
- SVM with linear kernel

Decision tree based system:
- Decision trees on phoneme lattices
- Normalized class count super vectors (38527 features/vector)
- SVM with linear kernel

	EER [%]
posterior bigrams	12,52
decision tree counts	13,23

All results on SRE2006 1conv4w-1conv4w English only

Factor analysis - flavors

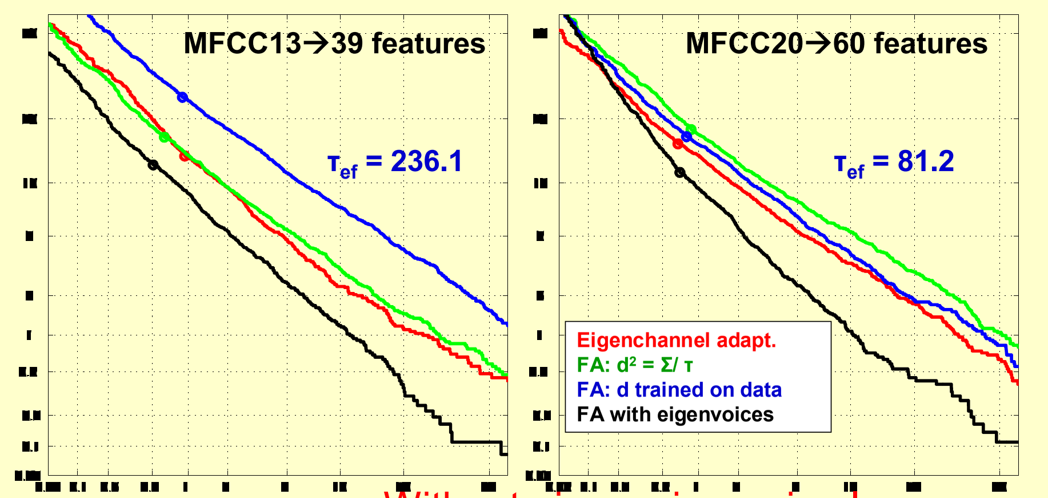


$$M = m + vy + dz + ux$$

All hyperparameters can be trained from data using EM

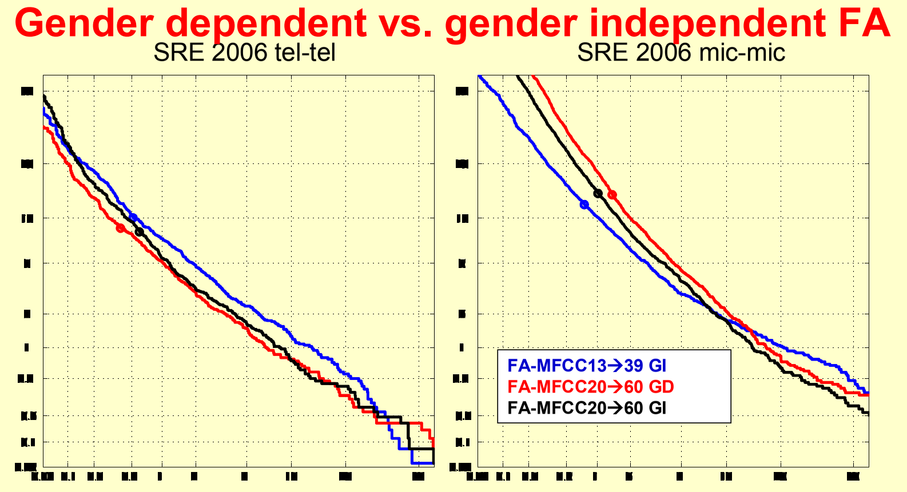
- **Relevance MAP adaptation**
  $M = m + dz$ with $d^2 = \Sigma/\tau$
- **Eigenchannel adaptation** (SDV, BUT)
  - Relevance MAP for enrolling speaker model
  - Adapt spk. model to test utt. using eigchannels estimated by PCA
- **FA without eigenvoices,** with $d^2 = \Sigma/\tau$ (QUT, LIA)
- **FA without eigenvoices,** with **d** trained from data (CRIM)
  can be seen as training different τ for each supervector coefficient
  Effective relevance factor τ_{ef}= trace(Σ)/ trace(d^2)
- **FA with eigenvoices** (CRIM)

SRE 2006 (all trials, det1)



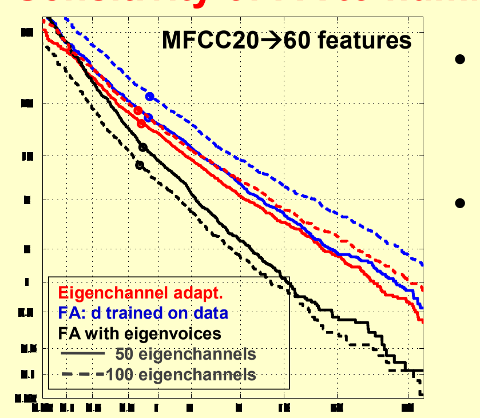
- Without eigenvoices, simple eigenchan. adaptation seems to be more robust than FA.
- FA with trained **d** fails for MFCC13→ 39 features. Too high τef? Caused by HLDA?
- FA with eigenvoices significantly outperforms the other FA configurations.

Gender dependent vs. gender independent FA



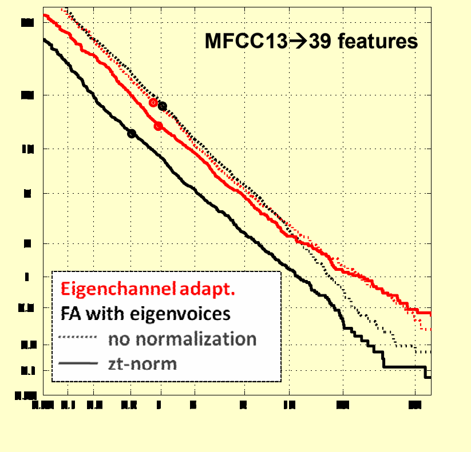
- Halving the number of parameters of FA-MFCC20→60 system by making it gender indep. degrades the performance on telephone and improves on microphone condition

Sensitivity of FA to number of eigenchannels

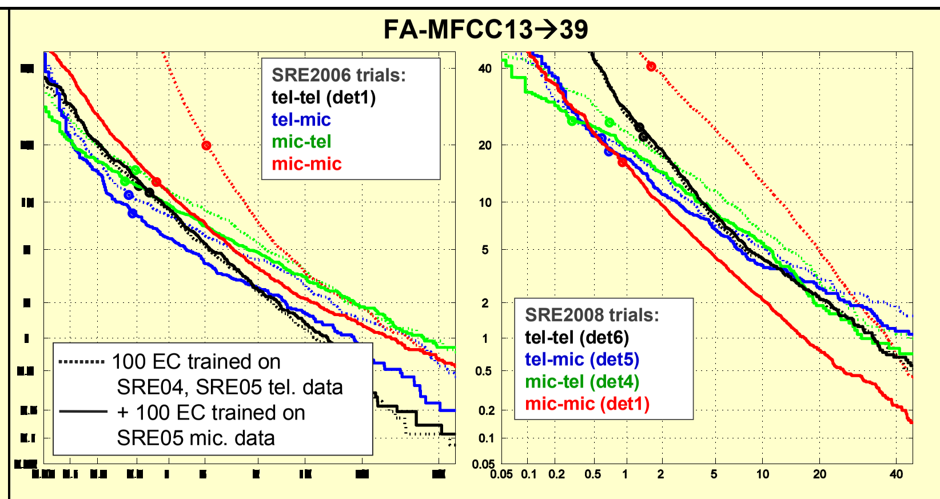


- FA systems without eigenvoices seem not to be able to robustly estimate increased number of eigenchannels
- However, we benefit from more eigenchannels significantly after explaining the speaker variability by eigenvoices

Importance of zt-norm

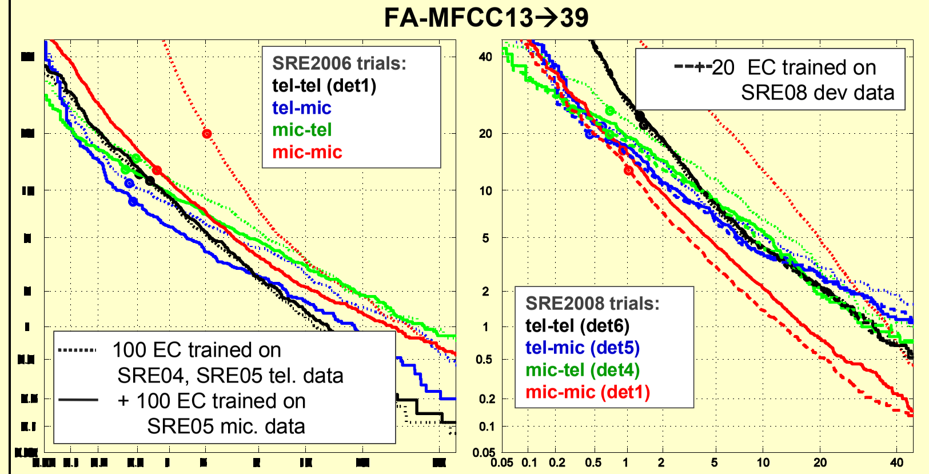


- zt-norm is essential for good performance FA with eigenvoices.



Training eigenchannels for different channel conditions

Negligible degradation on tel-tel condition and huge improvement particularly on mic-mic condition is obtained after adding eigenchannels trained on microphone data to those trained on telephone data.



Training additional eigenchannels on SRE08 dev data

Significant improvement is obtained on microphone condition for eval data after adding eigenchannels trained on SER08 dev data (all spontaneous speech from the 6 interviewees).

References

[Mason2005] M. Mason et al: Data-Driven Clustering for Blind Feature Mapping in SpkID, Eurospeech 2005.

[Chang2001] C. Chang et al.: LIBSVM: a library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm

[Hain2005] T. Hain et al.: The 2005 AMI system for RTS, Meeting Recognition Evaluation Workshop, Edinburgh, July 2005.

[Stolcke2005/6] A. Stolcke: MLLR Transforms as Features in SpkID, Eurospeech 2005, Odyssey 2006

[Brummer2004] N. Brummer: SDV NIST SRE'04 System description, 2004.

[Brummer_FoCal] N. Brummer: FoCal: Toolkit for fusion and Calibration, www.dsp.sun.ac.za/~nbrummer/focal

[Campbell2006] W. M. Campbell et al., "SVM Based Speaker Verification Using a GMM Supervector and NAP Variability Compensation," ICASSP 2006.