Detecting sleepiness by fusing classifiers trained with novel acoustic features

Multimodal Signal Processing (MSP)
Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas

Tauhidur Rahman
Soroosh Mariooryad
Shalini Keshavamurthy
Gang Liu
John H.L. Hansen
Carlos Busso*

Introduction

About Sleepiness

- It results from mental or physical fatigue, strain and exhaustion
- It impairs cognitive abilities, reducing the efficiency to perform operationally relevant tasks

Why is Sleepiness Detection?

- In-car technologies to prevent car accidents
  - 22-24% of car accidents occur due to sleepy drivers [Klauer et al., 2006]
- Studying sleep disorders
- Designing Human-Machine Interfaces

Approach

Goal

- Detecting sleepiness from speech
  - It can be captured from nonintrusive sensors

Approach

- Evaluate novel acoustic features for detecting sleepiness
  - Likelihoods from reference neutral models
  - Statistics of F0 contour across voiced segments
  - PMVDR+SDC
- Decision level fusion of individual classifiers
  - INTERSPEECH 2009 Emotion Challenge (41.7% -> 44.0%) [Schuller et al., 2011b]

Database

This study uses Sleepy Language Corpus (SLC)

- 21 hours of speech from 99 participants (9089 turns)
- Recordings in a realistic car-environment or in a lecture room
  - Isolated vowels
  - Read speech
  - Commands/requests
  - Spontaneous speech
- Karolinska sleepiness scale (1 -extremely alert to 10 - extremely sleepy)
  - Above 7.5 is considered sleepy (SL)
- Divided speaker independently into three groups:
  - Training (~40%), Development (~30%), Testing (~30%)

Classifiers for Sleepiness Detection

Baseline SVM System (λ_B)

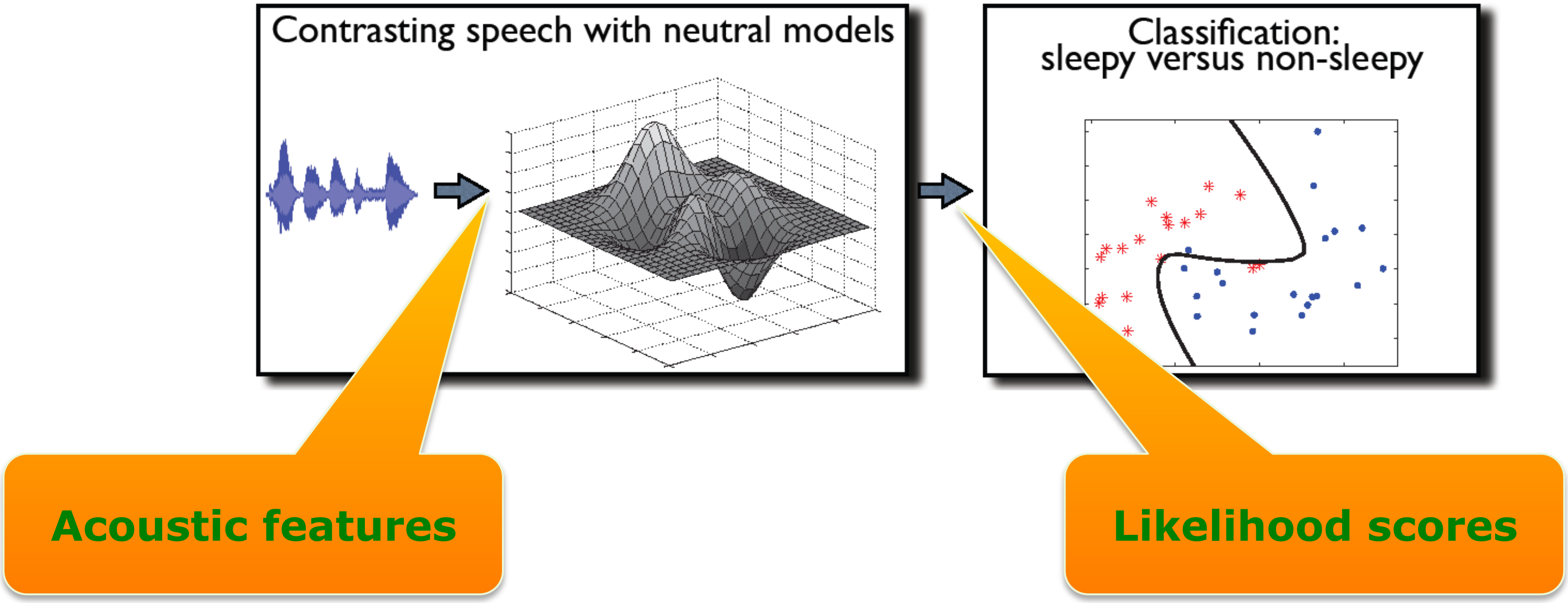
- A baseline classifier is trained as reference [Schuller et al., 2011a]
- Linear kernel support vector machine (SVM) with sequential minimal optimization (SMO)
- INTERSPEECH 2011 feature set:
  - 59 Low-Level Descriptors
  - 33 base functionals and 6 F0 functionals
  - Altogether: 4368 sentence level features
- Synthetic Minority Oversampling Technique (SMOTE)

INTERSPEECH 2011 FLORENCE

Classifiers for Sleepiness Detection

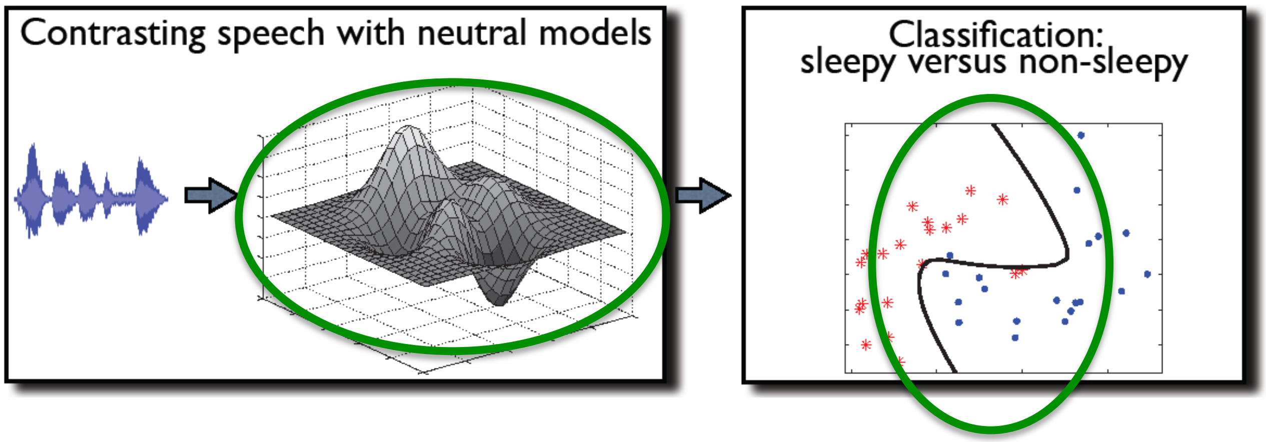
Likelihoods from reference neutral models (λ_L)

- Quantify deviations from neutral speech [Busso et al., 2007,2009]
  - Train models with neutral speech
  - Use likelihoods (fitness measure) as features for classification



Classifiers for Sleepiness Detection

Likelihoods from reference neutral models (λ_L)

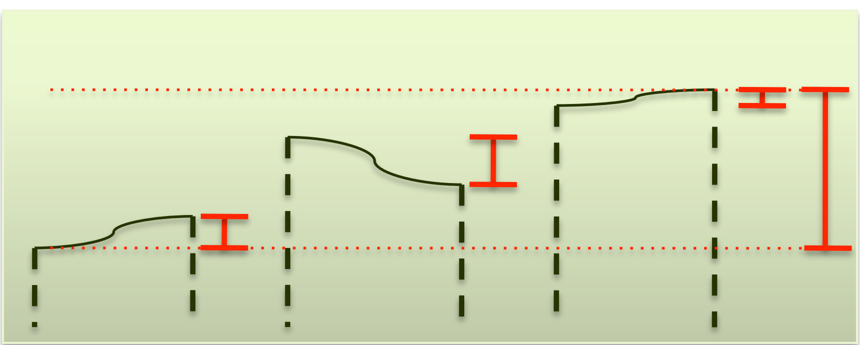


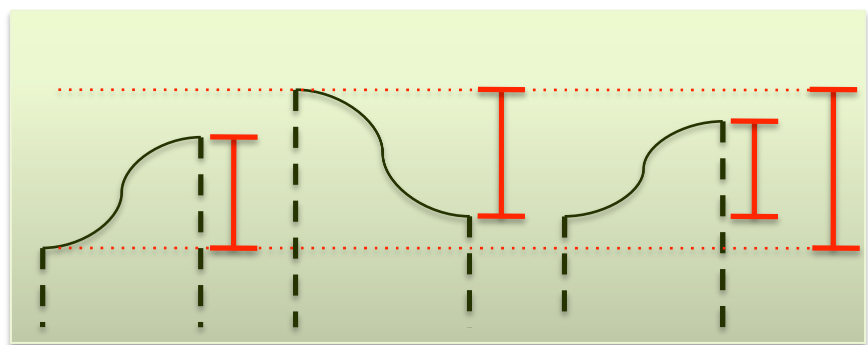
- Neutral Models based on Gaussian Mixture Models (GMMs)
  - 4 mixtures for each of the 4368 sentence level features
  - Wall Street Journal-based corpus
- Linear Kernel SVM trained on Likelihoods of the models

Classifiers for Sleepiness Detection

System based on voiced segment statistics (λ_F)

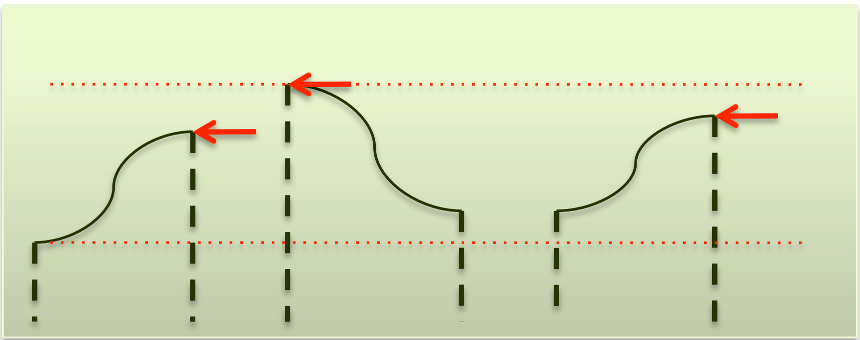
- Basic functionals such as range, maximum, quartiles, slope, curvatures are estimated over voiced segmented regions. [Busso et al, 2009]
- For each sentence, functionals are estimated again across these voiced segment statistics
- Provide insights about local dynamics of the pitch contour
- Linear Kernel SVM

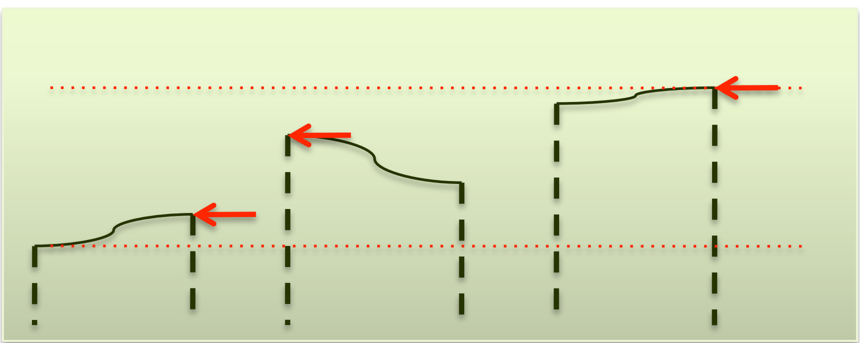




Example, mean of voiced segments ranges

System based on voiced segment statistics (λ_F)





STATISTICS OF THE VOICED REGION
Mean of the voiced segment ranges
Mean of the voiced segment maximums
Mean of the voiced segment minimums
Mean of the voiced segment lower quartiles
Mean of the voiced segment upper quartiles
Mean of the voiced segment interquartile ranges
Mean of the voiced segment slopes
Mean of the voiced segment curvatures
Mean of the voiced segment inflexions
Max. of the voiced segment slopes
Max. of the voiced segment curvatures
Max. of the voiced segment inflexion
Max. of the voiced segment mean
Std. of the voiced segment means
Std. of the voiced segment slopes
Std. of the voiced segment curvatures
Std of the voiced segment inflexions

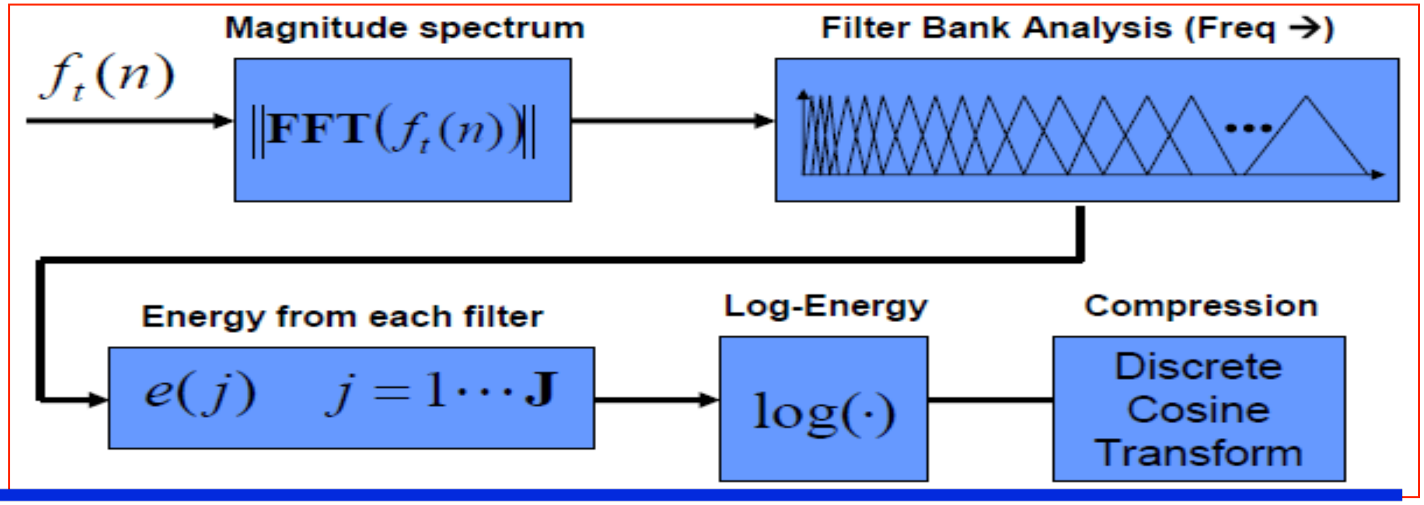
Classifiers for Sleepiness Detection

GMM trained with PMVDR & SDC (λ_p)

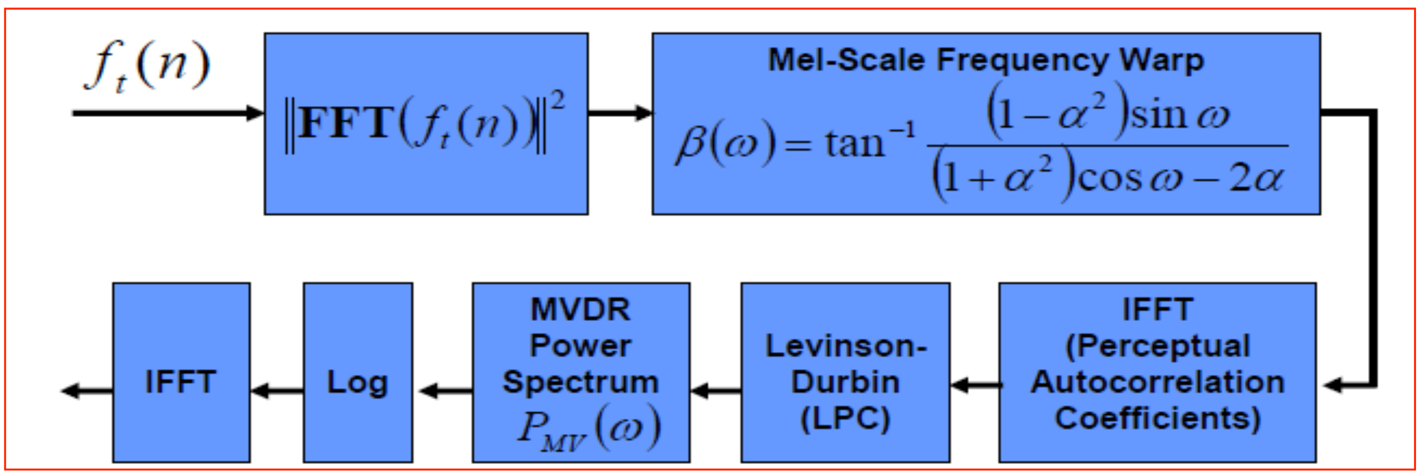
- Perceptual Minimum Variance Distortionless Response (PMVDR)
  - PMVDR is able to better model the upper spectral envelope than MFCC
  - Robust against noise
  - 10-dimensional PMVDR feature vector
- Shifted Delta Cepstrum (SDC)
  - SDC incorporates additional temporal information

MFCC vs. PMVDR

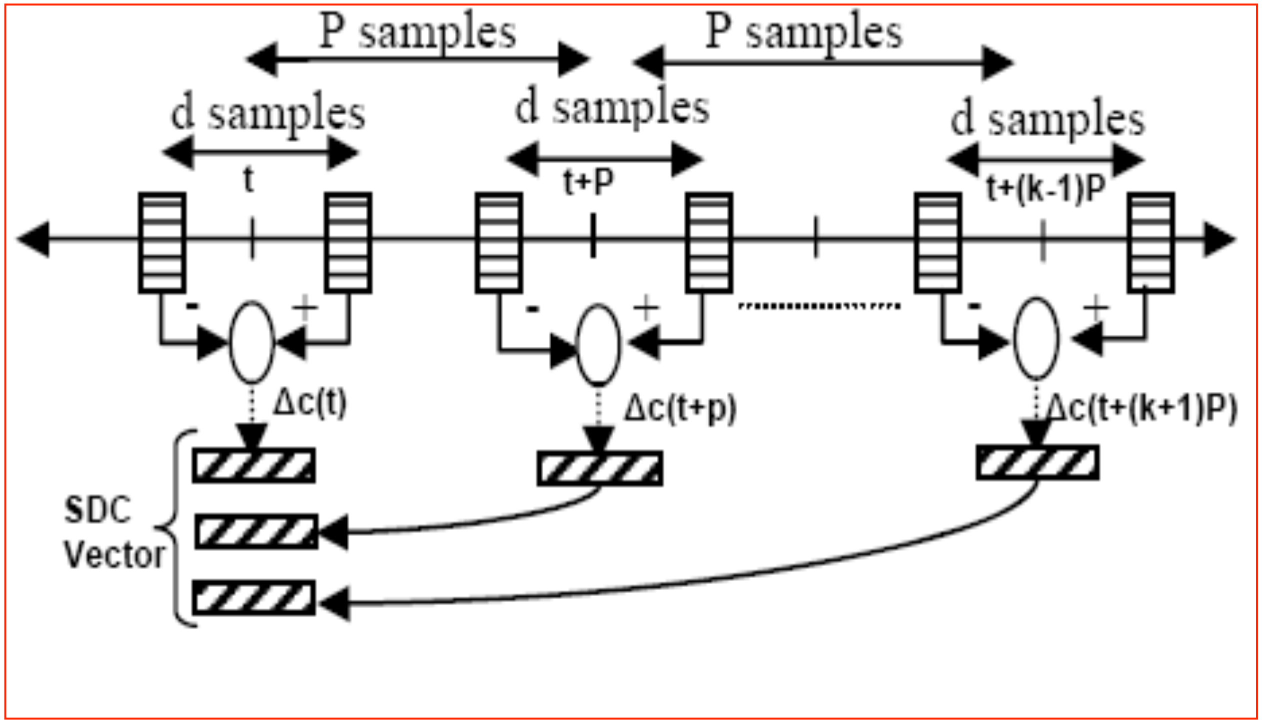
MFCC



PMVDR



Shifted delta cepstrum (SDC) N-d-P-K



11-1-3-3

Classifiers for Sleepiness Detection

- A GMM is trained frame by frame (PMVDR & SDC)
  - Normalized sum of the likelihoods at the sentence level

GMM trained with MFCCs (λ_M)

- 12 MFCC coefficients and their delta and delta-delta
- GMM is trained

Performance individual classifiers

Performance of Sleepiness Detection Systems – on the Development set:

Classifier	%WA	%UA	% Recall	%Precision	Class
λ_B Baseline	70.7	67.5	80.1	75.1	NSL
			54.8	58.1	SL
λ_L Likelihoods	66.7	64.0	74.0	73.2	NSL
			53.9	54.9	SL
λ_F VS-statistics	50.6	57.5	31.1	76.6	NSL
			83.9	41.7	SL
λ_P PMVDR & SDC	59.5	59.3	67.1	70.3	NSL
			51.8	48.1	SL
λ_M MFCC	61.4	57.3	64.9	68.7	NSL
			49.8	45.4	SL

Motivation for fusion

$$Q_{ij} = \frac{N_{11}N_{00} - N_{01}N_{10}}{N_{11}N_{00} + N_{01}N_{10}}$$

Classifier	λ_L Likelihoods	λ_B Baseline	λ_F VS-statistics	λ_P PMVDR & SDC	λ_M MFCC
λ_L Likelihoods	1.000	0.8689	0.2451	0.3705	0.7911
λ_B Baseline		1.000	0.1854	0.4904	0.8272
λ_F VS-statistics			1.000	0.0322	0.4754
λ_P PMVDR & SDC				1.000	0.4360
λ_M MFCC					1.000

- Classifiers with Different Characteristics

Fusion

Classifier	%WA	%UA
λ_B Baseline	70.7	67.5

Feature Level

- 7812 features
  - Baseline features
  - Likelihood features
  - F0 statistics
- Linear kernel SVN with with Sequential Minimal Optimization (SMO)
- Chi-squared feature selection technique

Features #	%WA	%UA
7812(all feature)	66.20	65.25
5000	68.20	66.90
3000	68.00	67.40
1000	63.30	64.00

Fusion

Decision Level

- Maximum Likelihood Decision

$$\hat{\omega} = \underset{\omega_\theta}{\operatorname{argmax}} \; P(\omega_{\lambda_1}, \omega_{\lambda_2}, \ldots, \omega_{\lambda_n} | \omega = \omega_\theta)$$
$$= \underset{\omega_\theta}{\operatorname{argmax}} \; \frac{P(\omega_{\lambda_1}, \omega_{\lambda_2}, \ldots, \omega_{\lambda_n}, \omega = \omega_\theta)}{P(\omega = \omega_\theta)}$$

  - Hard Decisions

$$\omega_{\lambda_i} \in \{SL, NSL\}$$

  - Soft Decisions

$$\omega_{\lambda_i} = P(\omega = SL)$$

Fusion: Results

Classifier	%WA	%UA
λ_B Baseline	70.7	67.5

Decision Level

- Train Set
  - Training the Models
- Development Set
  - 90% Training Fusion
  - 10% Test
  - 10-fold Cross Validation

Classifier	Hard fusion %WA	Hard fusion %UA	Soft fusion %WA	Soft fusion %UA
λ_B, λ_L	69.3	67.5	65.8	64.1
$\lambda_B, \lambda_L, \lambda_F$	68.9	68.2	67.1	66.6
$\lambda_B, \lambda_L, \lambda_M$	69.5	68.1	69.1	67.8
$\lambda_B, \lambda_L, \lambda_P$	70.4	68.3	68.9	68.2
$\lambda_B, \lambda_M, \lambda_P$	69.5	67.0	70.6	68.3
$\lambda_B, \lambda_L, \lambda_F, \lambda_P$	70.1	68.1	67.9	67.2
$\lambda_B, \lambda_L, \lambda_M, \lambda_P$	70.1	68.2	70.2	68.7
$\lambda_B, \lambda_L, \lambda_F, \lambda_M, \lambda_P$	68.8	67.1	68.9	67.7

λ_L Likelihoods
λ_B Baseline
λ_F VS-statistics
λ_P PMVDR & SDC
λ_M MFCC

Fusion: Results

Decision Level

- Train Set
  - Training the Models
- Development Set
  - Estimate conditional probabilities (soft decision)
- Test Set
  - Evaluation

	Schuller et al., 2011a [%]	$\lambda_B, \lambda_L, \lambda_M, \lambda_p$ [%]	Δ [%]
%WA	73.0	74.1	+1.1
%UA	70.3	71.0	+0.7

Conclusions

- Sleepiness can be detected using acoustic feature
- Feature level fusion was not as effective as late fusion
- Late-fusion strategies on hard and soft decisions are opted to improve the accuracy of individual classifiers
- Best Performance on testing set: 71.0% (UA)

Thanks!

Questions?

References

- S.Klauer, T.Dingus, V.Neale, J.Sudweeks, and D.Ramsey, “The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data,” National High- way Traffic Safety Administration, Tech. Rep., 2006.
- B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, “The INTERSPEECH 2011 speaker state challenge,” in Inter- speech 2011.Florence, Italy: ISCA, August 2011.
- B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge” Speech Communication 53 (9-10): 1062-1087 (2011).
- C. Busso, S. Lee, and S. Narayanan, “Analysis of emotionally salient aspects of fundamental frequency for emotion detection,” IEEE Transactions on Audio, Speech and Language Processing, vol. 17, no. 4, pp. 582–596, May 2009.
- C. Busso, S. Lee,and S. Narayanan,“Using neutral speech models for emotional speech analysis,” in Interspeech 2007 - Eurospeech, Antwerp, Belgium, August 2007, pp. 2225–2228.

Q statistic

$$Q_{ij} = \frac{N_{11}N_{00} - N_{01}N_{10}}{N_{11}N_{00} + N_{01}N_{10}}$$

N_{11}: is the number of both classifiers making the correct classification

N_{10}: is the number of y_i being correct and being y_j incorrect

N_{01}: is the number of y_i being incorrect and being y_j correct

N_{00}: is the number of both classifiers making the incorrect classification