Foreign accent recognition project

Ville Hautamäki
School of Computing, University of Eastern Finland (UEF)

Funding by MATINE: 67 340 EUR

Speech @ UEF (1/2)

Project PI: **Dr. Ville Hautamäki**:

Had Academy of Finland post-doc project on the same topic.

PhD Students:

Hamid Behravan Foreign accent recognition backend classifier work.

Ivan Kukanov Deep learning (convolutive neural networks, etc)

Msc Student:

Trung Ngo Trong Deep learning (long-short term memory, etc)

Bsc Students:

Anssi Kanervisto & Hannu Sillanpää

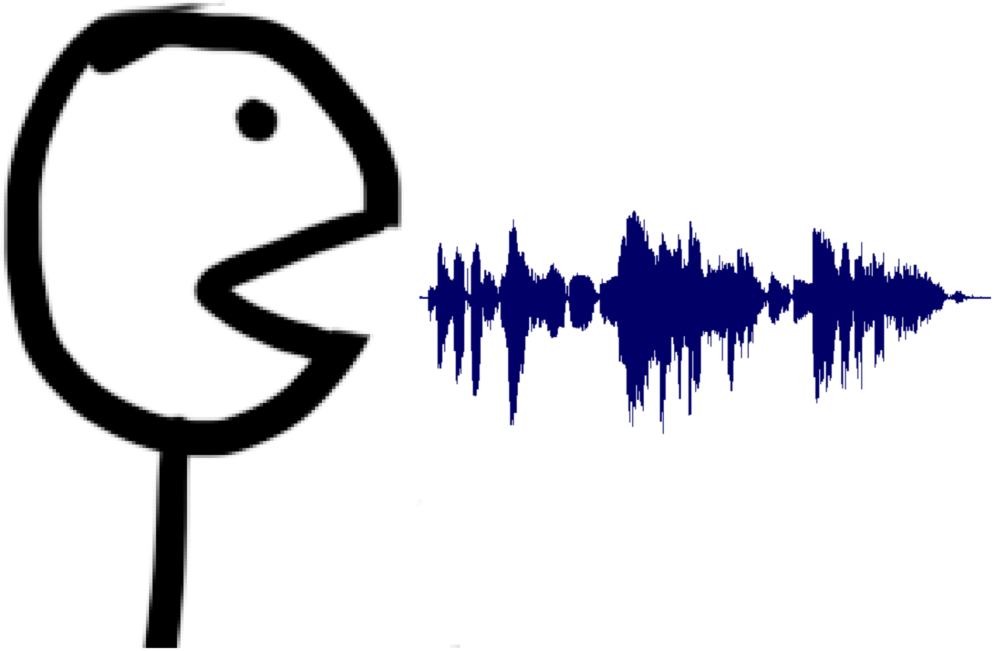
Speech @ UEF (2/2)

- PI Dr. Tomi Kinnunen + 1 post-doc + 3 PhD students.
- Speaker recognition projects
  - Access control and anti-spoofing

    OCTAVE H2020 project (innovation action)
  - Audio-visual anti-spoofing biometric recogntion project

    BIOSECURE, Academy of Finland funded

Soft-biometrics from speech signal



We recognize some speaker characteristics:

- Age
- Language
- Accent
- Identity
- Gender

Identity verification in border control

Can we verify the passport identity without the chip and biometric reader?

Soft biometrics is our approach. In passport we have written information about country of origin, age etc.

Foreign accent in border control

The proposed system should be assistive to border control officers. If passport says that traveler mother tongue is French but the system says that it is unlikely true then it should alert the officer.

Research questions

Can we improve recognition accuracy by the use of deep learning?

Note that this task is inherently difficult!

Would the deployed system be useful in practice?

How to evaluate the system performance?

Unknown test utterance is attached with a *claim*, such as foreign accent is Finnish.

Then we have a hypothesis H_1 that the claim is correct and an anti-hypothesis H_0 that the claim is incorrect.

System can then make two types of errors: *miss* and *false alarm* (or Type I and Type II errors).

Equal error rate is the operating point where these error rates are equal.

Universal Speech Attributes

Manner of articulation describes how the tongue, lips, jaw, and other speech organs are involved in making a sound make contact:

Nasal: */m/ , /n/*

Stop: /p/, /b/, /k/

Fricative: /f/, /v/, /θ/

Vowel: /a/, /e/, /i/

Approximant: /w/, /r/

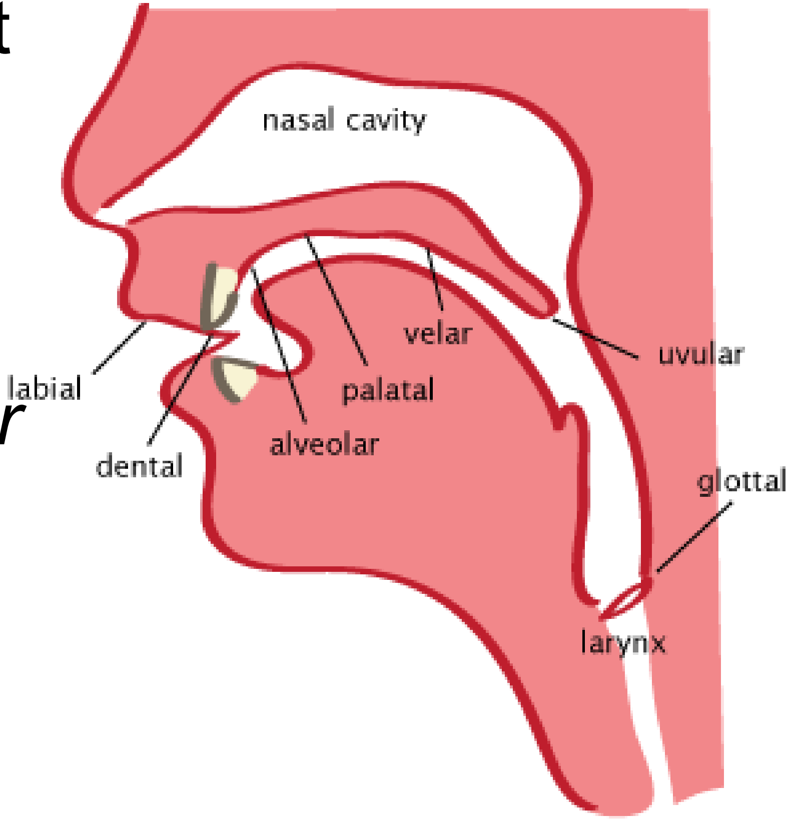
Glide: /j/

And

Voicing: Voicing refers to the presence or absence of vocal fold vibration.

Place of articulation
specifies the position at which a constriction in the vocal tract occurs.

Example : /d/
Place of articulation = alveolar
Manner of articulation = stop
Voicing = voiced (The vocal folds are vibrating.)



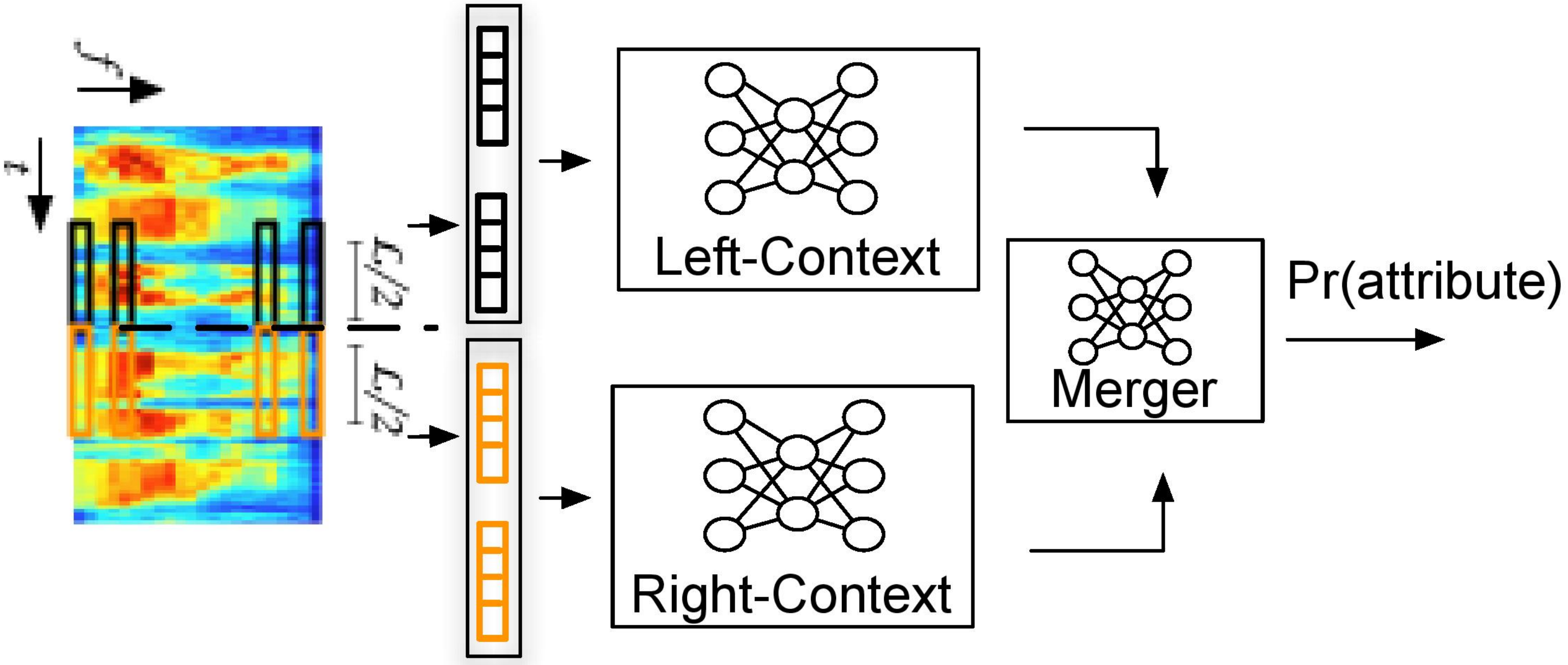
Why Speech Attributes?

- Those are language universal speech descriptors. All speech in all languages can be described using speech attributes.

- Statistics of their co-occurrences change from one language to another.

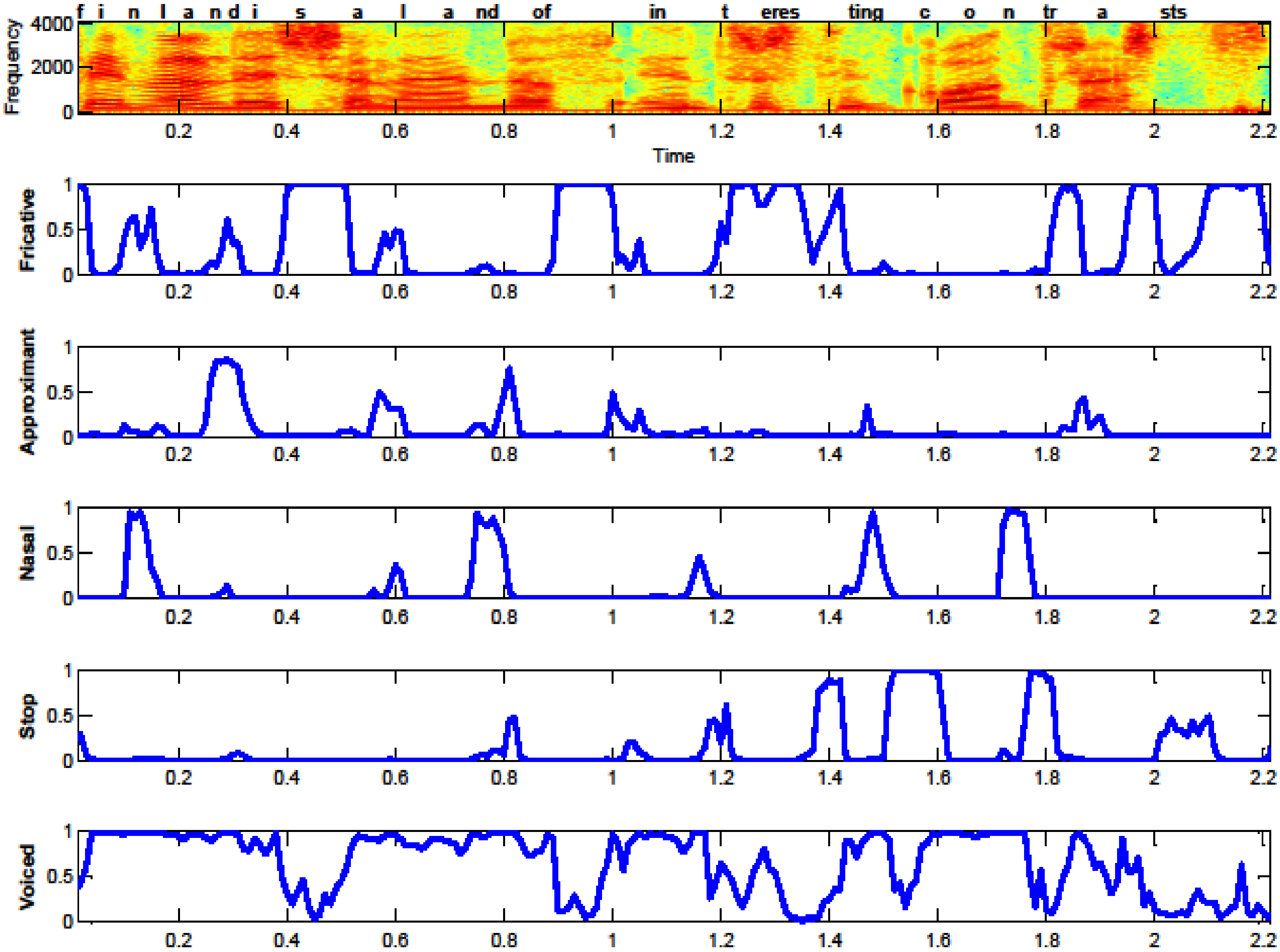
- Universal attribute detectors can be designed by sharing data among different languages.

How Attributes Are Extracted From Audio Signals?

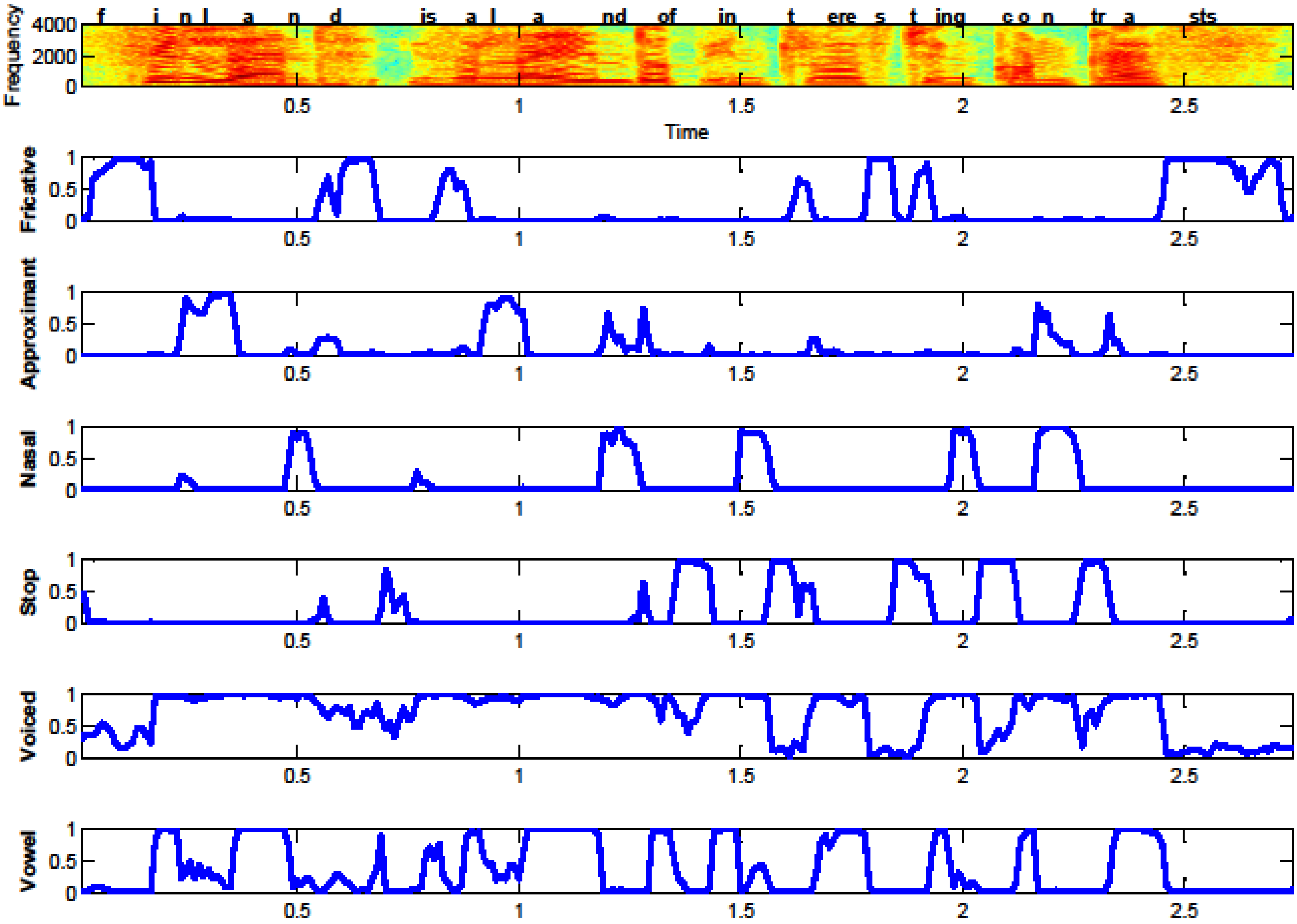


The internal structure of an attribute detector

Example: Finnish speaker



Example: Indian speaker

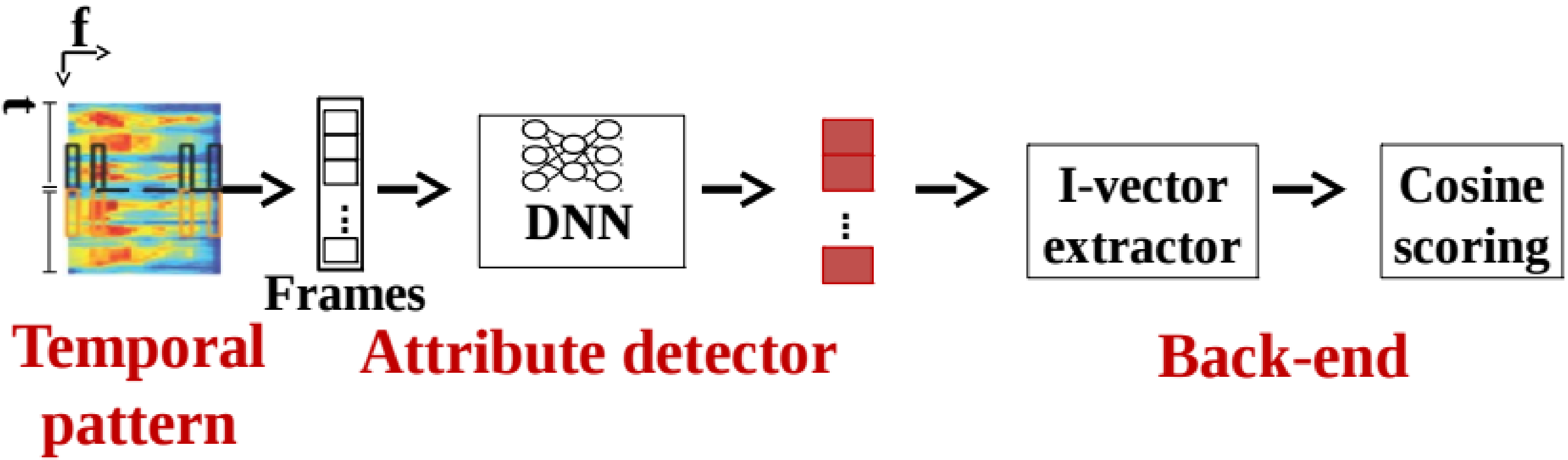


Rationale for the deep neural architechture

- In previous model, detectors were independent.
- Now correlations between detectors can be modeled.
- Hope was to improve the accuracy of place of articulation by adding depth to the network.
- Detection of foreign accents is based on the idea that speakers will mispronounce words. This mispronunciation is can be seen in place of articulation, such as /a/ and /ä/, where the difference is in the place only.

The proposed DNN architechture

We train one network for manner and place separately. Output layer is set to be softmax, this poses a difficulty as it assumes that only one attribute is active in any given time.



DNN experiments with OGI-TS corpus

Corpus is a multilingual telephone quality speech with transcriptions. Languages are English, German, Hindi, Japanese, Mandarin and Spanish.

Corpus is divided into three disjoint portions: *training* (train network), *cv* (used in avoiding overfit) and *testing* (compute error rates).

Using the phonetical theory, the mapping from phonemes to attributes is performed. These attribute labels are used in training, cv and testing.

Results of DNN experiments (1/2)

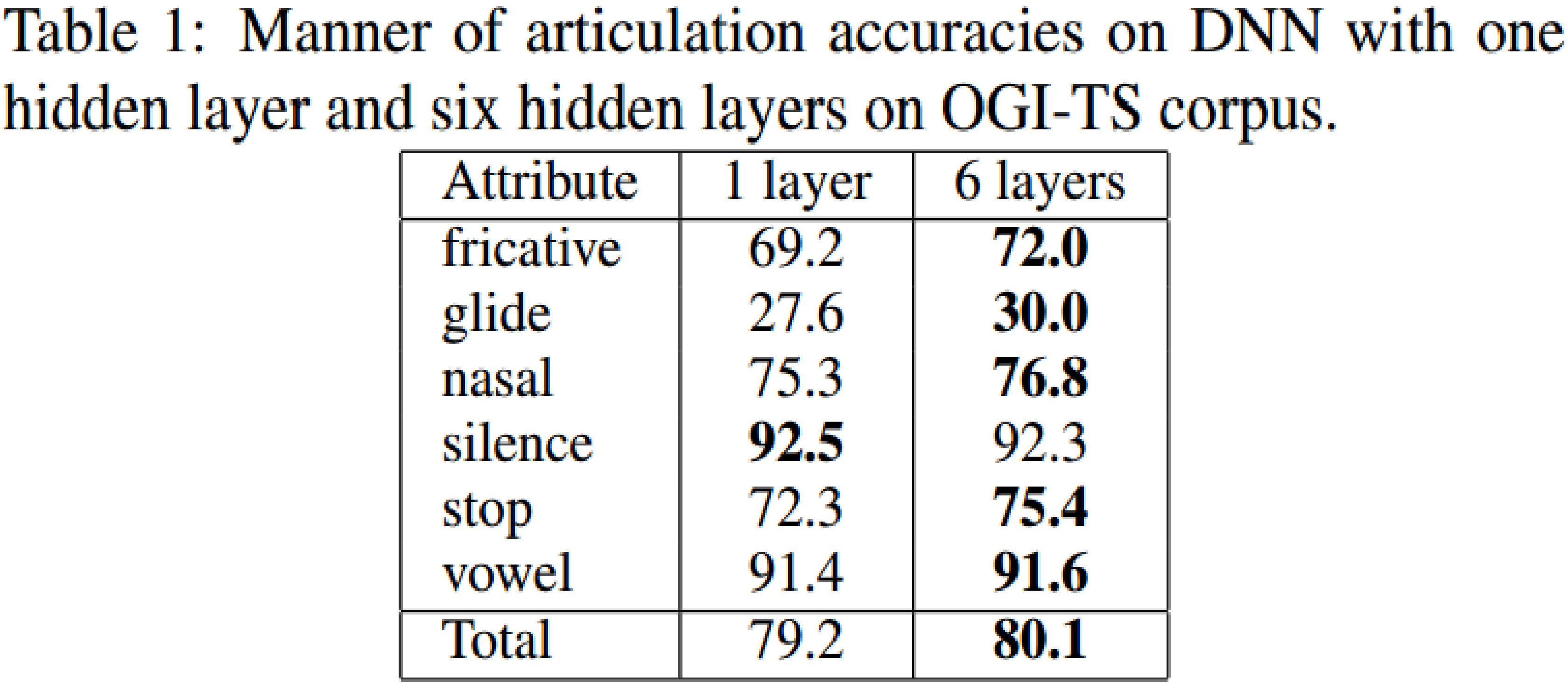


Table 1: Manner of articulation accuracies on DNN with one hidden layer and six hidden layers on OGI-TS corpus.

Attribute	1 layer	6 layers
fricative	69.2	**72.0**
glide	27.6	**30.0**
nasal	75.3	**76.8**
silence	**92.5**	92.3
stop	72.3	**75.4**
vowel	91.4	**91.6**
Total	79.2	**80.1**

Results of DNN experiments (2/2)

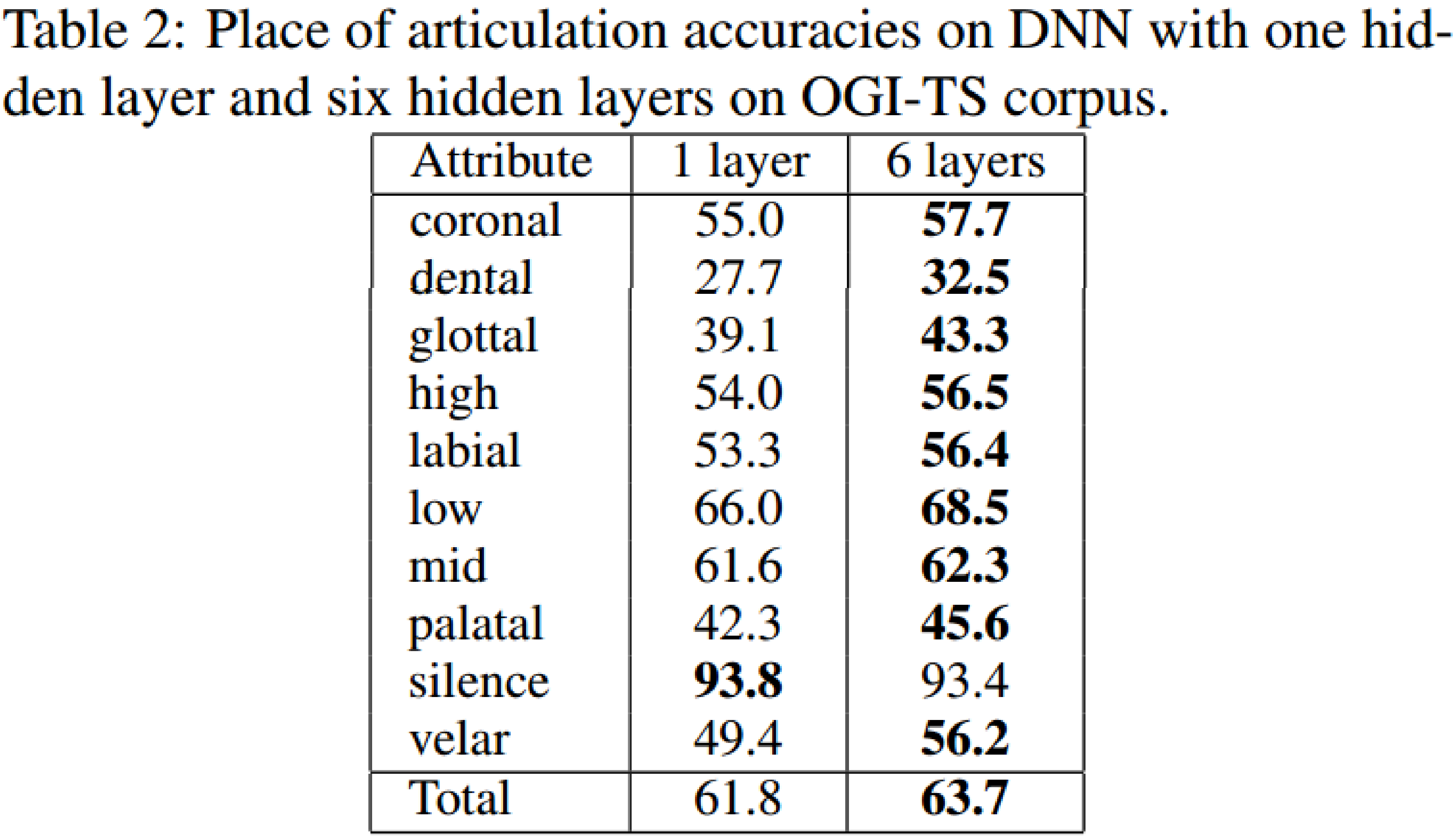


Table 2: Place of articulation accuracies on DNN with one hidden layer and six hidden layers on OGI-TS corpus.

Attribute	1 layer	6 layers
coronal	55.0	**57.7**
dental	27.7	**32.5**
glottal	39.1	**43.3**
high	54.0	**56.5**
labial	53.3	**56.4**
low	66.0	**68.5**
mid	61.6	**62.3**
palatal	42.3	**45.6**
silence	**93.8**	93.4
velar	49.4	**56.2**
Total	61.8	**63.7**

English as a foreign language corpus

We constructed this corpus from NIST speaker recognition evaluation (SRE) 2008 telephone corpus. It has large number of speakers and recorded utterances, where many spoke English as a second language.

In our experiments we used seven different foreign accents, with 511 different speakers and 1262 different utterances. Set was divided into training and testing portions, in such a way that speaker is either in the test or the training portion.

Foreign accent recognition results

Using DNN's we can push the average error rate to close to 10% It is expected that using both DNN place and DNN manner in together will improve performance.

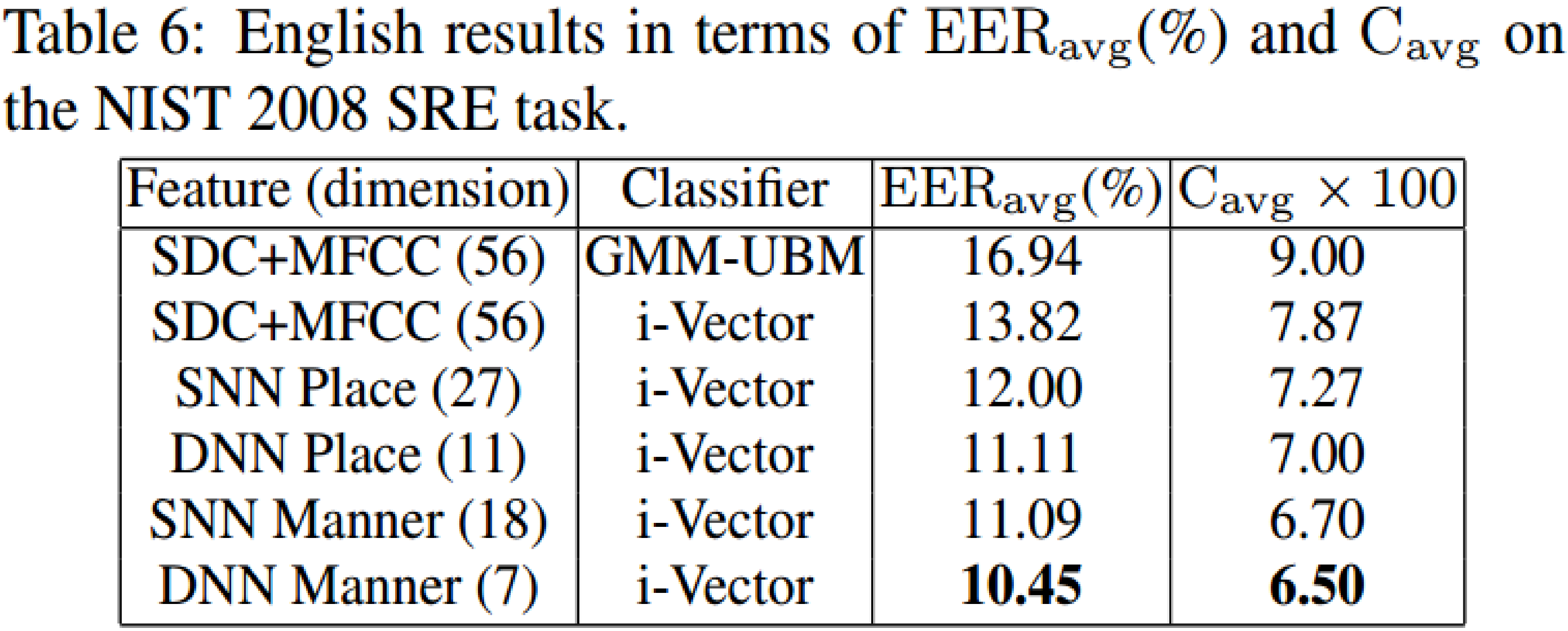


Table 6: English results in terms of $\mathrm{EER}_{\mathrm{avg}}(\%)$ and $\mathrm{C}_{\mathrm{avg}}$ on the NIST 2008 SRE task.

Feature (dimension)	Classifier	$\mathrm{EER}_{\mathrm{avg}}(\%)$	$\mathrm{C}_{\mathrm{avg}} \times 100$
SDC+MFCC (56)	GMM-UBM	16.94	9.00
SDC+MFCC (56)	i-Vector	13.82	7.87
SNN Place (27)	i-Vector	12.00	7.27
DNN Place (11)	i-Vector	11.11	7.00
SNN Manner (18)	i-Vector	11.09	6.70
DNN Manner (7)	i-Vector	**10.45**	**6.50**

Looking at per language results

We took a closer look at the per language results. Now to increase the amount of speech data in test portions we used jack-knifing with respect to speakers. We notice systematic improvement.

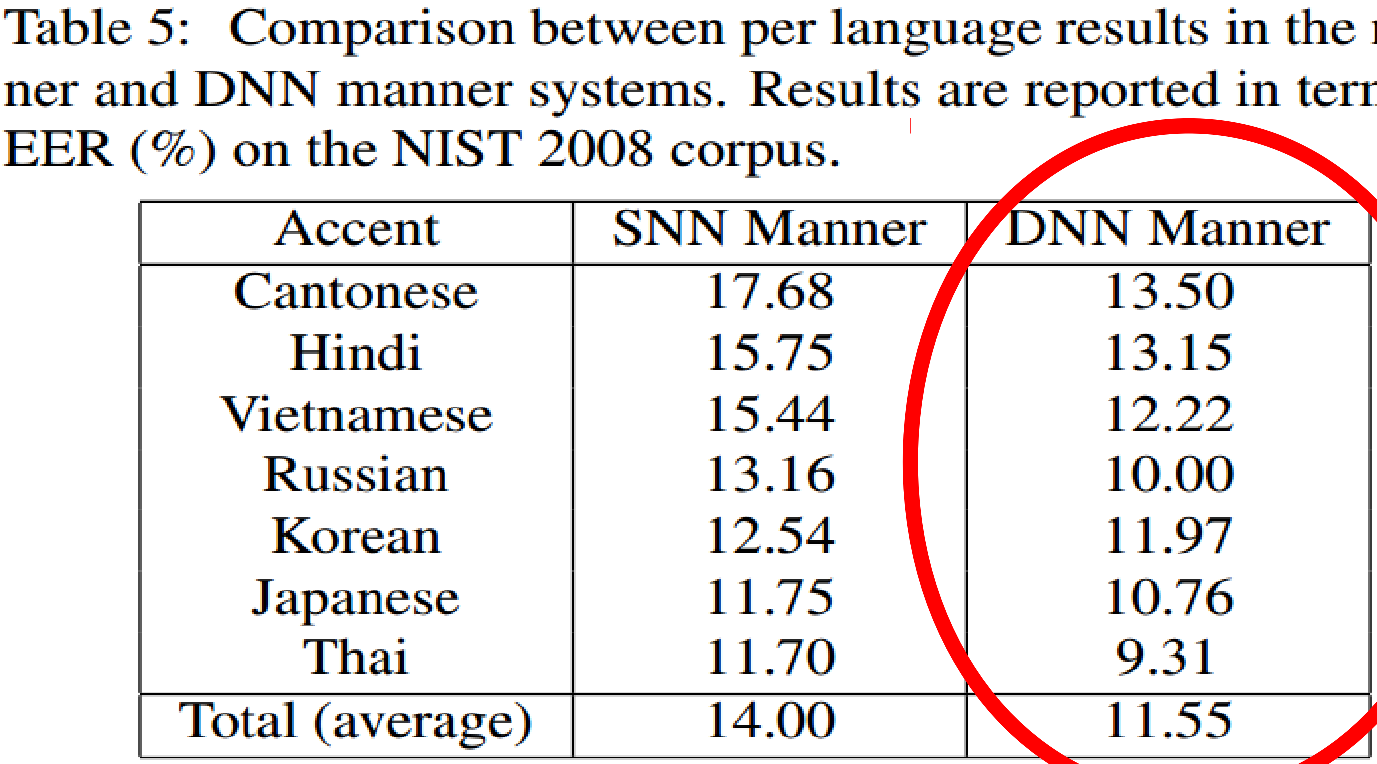


Table 5: Comparison between per language results in the manner and DNN manner systems. Results are reported in terms of EER (%) on the NIST 2008 corpus.

Accent	SNN Manner	DNN Manner
Cantonese	17.68	13.50
Hindi	15.75	13.15
Vietnamese	15.44	12.22
Russian	13.16	10.00
Korean	12.54	11.97
Japanese	11.75	10.76
Thai	11.70	9.31
Total (average)	14.00	11.55

Around 10% error.

Significance

- Lets assume this dataset generalizes to border control case:
  - 10 % of the travelers with fake passport are accepted and 10% of the genuine passport holders are rejected.
- We are working hard to further improve these results.
- What could be an acceptable error rate?

What was done

- Msc thesis of Ivan Kukanov on deep learning.
- Developed an Android demo app.
- Two accepted publications + two new papers in preparation.
- Msc thesis of Trung Ngo Trong in preparation.

  - Hamid Behravan, Ville Hautamäki, Sabato Siniscalchi, Tomi Kinnunen, and Chin-Hui Lee, "i-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition", *IEEE Transactions on Audio, Speech and Language Processing* (accepted).
  - Ville Hautamäki, Sabato Siniscalchi, Hamid Behravan, Valerio Mario Salerno and Ivan Kukanov, "Boosting Universal Speech Attributes Classification with Deep Neural Network for Foreign Accent Characterization", *Interspeech 2015*, Dresden, Germany, September 2015.

Extra task

- Since 2006 we, in UEF, have participated in the technology evaluations organized by NIST and funded by the US DoD. We receive only the data for free, but no funds. Evaluations are either language recognition or speaker recognition.
- This fall we had an excellent opportunity to participate in the NIST language recogntion 2015 evaluation.
- The project members (*Kukanov* and *Trong*) at UEF participated and the PI (*Hautamäki*) was the coordinator of the four participant consortium (Finland, Singapore, France and Italy).
- More information:

http://www.nist.gov/itl/iad/mig/lre15.cfm