NIST SRE 2008
UWS system

Benoît Fauve
John Mason

Swansea University
(formally University of Wales Swansea)

Speech and Image Research Group
Electronics Research Centre
School of Engineering, Swansea University, UK

http://eegalilee.swan.ac.uk
1) Why I am talking: “the most systems attempting the most conditions”. muscles/brawn

2) Other interesting point in UWS: simple single approach brains

Since SRE 04 (Roland Auckenthaler - very fast commercial system) we keep the tradition of submitting on all conditions with generic system (all Alize based) - a few adaptation to task types

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>10sec</th>
<th>short</th>
<th>long</th>
<th>summed</th>
</tr>
</thead>
<tbody>
<tr>
<td>10sec</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>short2</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3conv</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>8conv</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>long</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>3summed</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

✓: Main development
➔: extensions
Collaborative work

Used scores from LIA and IDIAP for 2 submissions
- IDIAP in UWS_1 10sec-10sec
- LIA+IDIAP in UWS_3 short2-short3

LIA:
- long term collaboration (joint PhD program is being set between UWS and LIA)
- Work on development of Alize (Driss Matrouf, Nicolas Scheffer, JF Bonastre), publications [1,2]

IDIAP:
- NIST new comer with original work

Looking for complete software: “Press a button and run” solution

Sorry ! ALIZE/SpkDet is not this perfect software !!!

But it includes a lot of what you need
  - Free/Open, more and more documented (…)
  - Feature stuff
  - GMM/UBM
  - SVM and supervector SVM
  - Factor Analysis and NAP
  - Evaluated on NIST-SRE (good level of performance)

ALIZE/LIA_SpkDet -> SpkDet -> Mistral (Seapker Id, Speaker segmentation, Multimodal processing, Demos…)
  - Mistral: funded by the French National Agency for Research (ANR).

Growing developer community: Subversion, Developers site, Wiki…
UWS development

UWS: main development was done over the last year [1,2,3]

Plan: setup scripts run systems, submit
-Find a script mixing test segments
- Some thresholds in a rush
  (look at our minDCF)

No development on interview segments: interesting 08 results on telephony speech results (det6/7/8, see later)

All background data from 04
Acoustic features only
(1) Short2-short3 \( (1\text{conv}4w-1\text{conv}4w) \) and other tasks with ‘long’ segments

<table>
<thead>
<tr>
<th></th>
<th>10sec</th>
<th>short</th>
<th>long</th>
<th>summed</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10sec</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>short2</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3conv</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8conv</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>long</td>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>3summed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1-1) One aspect of development:

Observed trends can be affected by a few files

SAD and/or empty *.ctm point to 20 ‘faulty’ files on 1conv4w-1conv4w task
- no conversation (person absent on the other side), line cut
- 1 noisy, 1 male labelled female (lucky dip, had empty ctm)

<table>
<thead>
<tr>
<th></th>
<th># tar</th>
<th># seg</th>
<th># trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>det1</td>
<td>635</td>
<td>2420</td>
<td>31243</td>
</tr>
<tr>
<td>det1'</td>
<td>628</td>
<td>2406</td>
<td>30591</td>
</tr>
<tr>
<td>det7</td>
<td>506</td>
<td>2122</td>
<td>20907</td>
</tr>
<tr>
<td>det7'</td>
<td>501</td>
<td>2111</td>
<td>20511</td>
</tr>
</tbody>
</table>

It concern ~2% of trials

We present two examples:
- Trends between channel compensation approaches
- Score normalisation
Irregular DETs on 05

**NIST05 female det1**

**GSL-NAP**
**SFA**
**GSL-FA**

*For references: GMM, GSL*

**GMM factor Analysis:**
**SFA**

**NIST05 male det7**

**SVM systems**

**GSL-NAP**: GMM Supervector Linear svm NAP

**SFA**: Factor Analysis on model and on features for test segments - traditional GMM scoring

**GSL-FA**: GSL on FA compensated models
GSL-NAP: GMM Supervector Linear svm NAP
SFA: Factor Analysis on model and on features for test segments - traditional GMM scoring
GSL-FA: GSL on FA compensated models
O5 filtered : score normalisation

SFA on all languages female 05

- $T_{norm}$
- $Z_{norm}$

Removing Faulty segments

At least one model from a cut file gives high scores
1-2) Which channel modelling?

Eigen channel is crucial, but which approach to use:

- **GSL-NAP**: GMM Supervector Linear svm NAP
- **SFA**: Factor Analysis on model and on features for test segments - traditional GMM scoring
- **GSL-FA**: GSL on FA compensated models
- Feature Level only…

References [1,2,7,8,9,10,14]
GSL-NAP or SFA? Difficult to find which system performs better (from 05 to 06, from male to female, from all lang to English)

GSL-FA seems to take the best of FA modelling and SVM paradigm
Complementary between channel compensation approaches?

GSL-FA seems to take the best of **FA modelling** and **SVM** paradigm
This is confirmed by looking at **fusion**

Fusion of **GSL-FA + GSL-NAP**, **GSL-FA + SFA** (not plotted here) **doesn’t help**

If there are **no complementarities between approaches** let’s look at **front-ends**...
1-3) MFCC+LFCC

- GSL-FA with LFCC - weightSAD - order50
- GSL-FA with MFCC - meanSAD - order49
- Fusion

- Late addition: only used on short2-short3

- Changed 3 elements but **key** change is from LFCC to MFCC

- ~15% improvement for both EER and minDCF

- For female: MFCC<LFCC (importance of higher resolution in high frequency)

![Graph showing EER and minDCF](image)
1-4) Score normalisation T/Z/ZT?

All languages All gender NIST 06 new key

No clear differences
-For English only difference between nonorm and Tnorm is larger
-Overall Tnorm only is good
To sum up

Usual approach: Accumulation of systems
To sum up

Usual approach:
Accumulation of systems

Our approach on short2-short3

UWS_1
1-5) Results on 08 det6

-FA-SGMM: LFCC - weightSAD - order50 UWS_2
--FA-SGMM: MFCC - meanSAD - order49
-Fusion UWS_1

minDCF: 0.367
actDCF: 0.402

-EER/DCF >> development set

-Good results for a ‘single’ system on det6/ det7/ det8

UWS_3: collaboration
UWS_MFCC + LIA_ztnorm + IDIAP_NAP
actDCF 0.446 - minDCF 0.397

Not as good as we hoped but we work on similar configurations/ background lists

Det6: tel-tel
Longer Training and other tasks

- **UWS_1**: (SVM)
  - LFCC wSAD → GSL-FA Tnorm → Fusion → UWS_1

- **UWS_2**: (SVM)
  - LFCC wSAD → GSL-FA Tnorm → UWS_2

- **UWS_3**: (GMM)
  - LFCC wSAD → SFA Tnorm → UWS_3
Longer Training

MFCC being a late addition (in UWS system) no time to use it on other conditions

**UWS_1**: GSL-FA we use 2 front-ends: only change is SAD
- **weightSAD / meanSAD** (# frame / different selection statistics)

Then we compare:
- **UWS_2**: GSL-FA (SVM)
- **UWS_3**: SFA (GMM)

(systems in Matrouf et al, Interspeech 07 [1] Fauve et al. TASLP [2])

![Graph showing performance comparison of different systems](image)
All ‘long’ durations

-UWS_4 (GSL-FA mSAD) **not submitted** but plotted because UWS_1 = UWS_2 + UWS_4

Observations:

- **Little difference** for conditions with **single speakers** (**combining both SADs** buys a little)

- With summed: \(\text{weighSAD} < \text{meanSAD}\) and \(\text{SFA} < \text{GSL-FA}\)

**For summed: Need to be checked with speaker segmentation**
(2) 10sec-10sec
and other tasks with ‘short’ segments

<table>
<thead>
<tr>
<th></th>
<th>10sec</th>
<th>short</th>
<th>long</th>
<th>summed</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10sec</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>short2</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>3conv</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8conv</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>long</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3summed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In 06 we submitted **GMM-UBM-MAP** systems with **different front-ends** (similar approach for USTC this year with reasonably good results)

This year it’s based on:
[3] *Fauve et al., “Improving the performance of text-independent short duration GMM- and SVM-based speaker verification” Odyssey 08*

**Very short duration issues:**
- Channel compensation approaches prove ineffective.
- Accurate frame selection is important (we use meanSAD)
- Limit of MAP approach (little # frames for stats estimation)

**1 front-end but different approaches** that prove complementary

- **G:** GMM-MAP
- **S:** GSL-ML (SV with relevance factor=0)
- **E** GMMEigenVoice

*And we add*
- **I:** SVM Max Kernel from IDIAP
- UWS_1: G+S+E+I
  actDCF: 0.877
  minDCF: 0.819

- UWS_2: G+S+E
  actDCF: 0.858
  minDCF: 0.844

**Eigenvoice**: (here full statistical learning not like in [3]) speakers from NIST’04.

**Disadvantage**
Better results can be obtained with larger eigenvoice space. Seems to require large number of speakers (1000-2000)

**Advantage**
good when dealing with data unmatched by large corpora (language, recording conditions, sampling freq)
2-2) 10sec testing

short2-10sec  8conv-10sec

Because 10s segments are involved
- We work with meanSAD features
- **Factor analysis** is used to build **model** (with eigenchannel compensation)
- **No compensation** on **10sec** test files: direct GMM scoring

Difference between UWS_1 UWS_2: **targeted Tnorm** (respectively 10s and 1conv segments) seems to **help on development set (06), not on 08**

Good result for 8conv-10sec.
the more data the better, the research is how to use it

Improvement over the last years for acoustic systems in ASV comes from efficient use of labelled per speaker background data from large corpora

• For eigenchannel
  - Learning on NIST’04 works well
  - Even mixing speakers (looking at intersession over an average of different speakers) it provides good results.

• For eigenvoice
  - The task is a bit more tricky
  - Need a lot more speakers (seems to be 1000s)
  - Digging into development data is not an easy task (specially w/o experience of past evaluations), different formats of keys from one year to another, same speakers appearing in 05/06 (extended) - 06/08 (10s)

Easier task for sites with access to full Switchboard and Fisher corpora

Suggestions for development data: provide a clear simplified key of speakers recordings over all NIST SREs
Conclusions

Future:
- Threshold / Calibration (over tasks)
- Speaker segmentation
- More generic system over tasks (eigenvoice, eigenchannel together)
- More data for eigenvoice

What worked for us:
- Selecting a single approach (SVM on FA compensated GMM)
- Simple system (no need of feat warping, no large cohort lists, no ZT/Znorm)
- Fusion MFCC+LFCC
Many thanks for your attention

Diolch!

Thank you! Merci!
References


