Phoneme and Sub-phoneme T-Normalization for Text-Dependent Speaker Recognition

Doroteo T. Toledano¹, Cristina Esteve-Elizalde¹, Joaquin Gonzalez-Rodriguez¹, Ruben Fernandez-Pozo² and Luis Hernandez Gomez²

¹ ATVS, Universidad Autonoma de Madrid, Spain
² GAPSS-SSR, Universidad Politecnica de Madrid, Spain

IEEE Odyssey 2008, Cape Town, South Africa, 21-24 Jan 08

Outline

1. Introduction
2. Text-dependent SR Based on Phonetic HMMs
   2.1. Enrollment and Verification Phases
   2.2. Experimental Framework (YOHO)
   2.3. Results with raw scores
3. T-Norm in Text-Dependent SR
   3.1. Plain (Utterance-level) T-Norm
   3.2. Phoneme-level T-Norm
   3.3. Subphoneme-level T-Norm
4. Results summary
5. Discussion
6. Conclusions
1. Introduction

- **Text-Independent Speaker Recognition**
  - Unknown linguistic content
  - Research driven by yearly NIST SRE evals

- **Text-Dependent Speaker Recognition**
  - Linguistic content of test utterance known by system
    - Password set by the user
      - Security based on password + speaker recognition
    - Text prompted by the system
      - Security based on speaker recognition only
  - No competitive evaluations by NIST
  - YOHO is one of the most extended databases for experimentation

- This work is on **text prompted systems with YOHO as test database**

2.1. Text-dependent SR based on phonetic HMMs: Enrollment Phase

- Speech parameterization (common to enrollment and test)
  - 25 ms Hamming windows with 10 ms window shift
  - 13 MFCCs + Deltas + Double Deltas → 39 coeffs

- Spk-indep, context-indep phonetic HMMs used as base models
  - 39 phones trained on TIMIT, 3 states left-to-right, 1-80 Gauss/state

- Spk-dep phonetic HMMs from transcribed enrollment audio
2.1. Text-dependent SR based on phonetic HMMs: Verification Phase

- Computation of acoustic scores for spk-dep and spk-indep models

- Acoustic scores → Verification score (\( \text{sc}_2(O, \lambda_D) \) removing silences)

\[
\text{sc}_1(O, \lambda_I) = \frac{1}{N} \left( \sum_{i=1}^{N} \text{acs}^D_i - \sum_{i=1}^{N} \text{acs}^I_i \right)
\]

\[
\text{sc}_2(O, \lambda_D) = \frac{1}{N_D} \sum_{i=1}^{k_D} \frac{k_D}{N_D} \sum_{j=1}^{p_{D,i}} \text{acs}^D_j - \frac{1}{N_D} \sum_{i=1}^{k_D} \frac{k_D}{N_D} \sum_{j=1}^{p_{D,i}} \text{acs}^I_j
\]

2.2. Experimental Framework (YOHO)

- YOHO database
  - 138 speakers (106 male, 32 female)
    - Enrollment data: 4 sessions x 24 utterances = 96 utterances
    - Test data: 10 sessions x 4 utterances = 40 utterances
  - Utterance = 3 digit pairs (i.e. “twelve thirty four fifty six”)

- Usage of YOHO in this work
  - Enrollment: 3 different conditions
    - 6 utterances from the 1st enrollment session
    - 24 utterances from the 1st enrollment session
    - 96 utterances from the 4 enrollment sessions
  - Test: always with a single utterance
    - Target trials: 40 test utterances for each speaker (138 x 40 = 5,520)
    - Non-tgt trials: 137 test utterances for each speaker (138 x 137 = 18,906)
  - One random utterance from the test data of each of the other users
2.3. Results with raw scores

- DET curves and %EERs with raw scores comparing
  - Baum-Welch Re-estimation vs. MLLR Adaptation
    - For optimum configuration of tuning parameters in each case (Gauss/state, regression classes, re-estimation passes)
  - Different amounts of enrollment material
    - 6, 24 or 96 utterances
- MLLR Adaptation provides better performance for all conditions
- Our baseline for this work is the curve for MLLR adaptation with 6 utterances

<table>
<thead>
<tr>
<th>Enrollment utterances (and sessions)</th>
<th>MLLR Adaptation</th>
<th>Baum-Welch Re-estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 (1 session)</td>
<td>4.6</td>
<td>5.6</td>
</tr>
<tr>
<td>24 (1 session)</td>
<td>2.1</td>
<td>3.2</td>
</tr>
<tr>
<td>96 (4 sessions)</td>
<td>0.9</td>
<td>1.9</td>
</tr>
</tbody>
</table>

3. T-Norm in Text-Dependent SR

- T-Norm in Text-Independent SR
  - Regularly applied with excellent results
  - Normalize each score w.r.t. distribution of non-target scores for
    - The same test segment
    - A cohort of impostor speaker models
- T-Norm in Text-Dependent SR
  - Rarely applied with only modest improvement
  - A few notable exceptions are
    - [M. Hébert and D. Boies, ICASSP’05], where T-Norm is the main focus and
    - [R.D. Zylca et al., Odyssey’04], where T-Norm is applied but is not the main focus
3.1. Plain (Utterance-level) T-Norm: Procedure

- Procedure in text-dependent SR is identical to T-Norm in text-independent SR
  - We call this Plain T-Norm or Utterance-level T-Norm to distinguish it from the other methods we propose

1. Compute verification scores for the same test utterance and a cohort of impostor speaker models:
   - Reserve a cohort of impostor speakers \{1, ..., M\}
   - Obtain MLLR speaker-adapted phonetic HMMs for those speakers
   - Compute verification scores for the same test utterance and those speaker models \{sc_1(O, \lambda_{D1}), sc_2(O, \lambda_{D2}), ..., sc_M(O, \lambda_{DM})\}

2. Normalize the verification score using the mean and standard deviation of the impostor scores obtained

\[
sc_2^{\text{Norm}}(O, \lambda_D) = \frac{sc_2(O, \lambda_D) - \mu_C}{\sigma_C}
\]

3.1. Plain (Utterance-level) T-Norm: Results (i)

- Plain (Utterance-level) T-Norm vs. No T-Norm on YOHO
  - Enrollment with only 6 utterances from 1 session and test with 1 utterance
  - 10 male and 10 female speakers reserved as cohort and not included in results
  - Cohort = 20 speaker models
  - MLLR adaptation

- Utterance-level T-Norm (Plain T-Norm) produces slightly worse results than doing nothing

- Perhaps due to very small cohort?
3.1. Plain (Utterance-level) T-Norm: Results (ii)

- Perhaps due to very small cohort?
- New experiment using a bigger cohort of models
  - But not speakers due to very limited amount of speakers in YOHO (32 f)
  - 4 speaker models by speaker in the cohort
  - Trained with the first 6 utterances in each session
- Slightly better results, but still the improvement achieved by T-Norm is very small
- Probably not only due to the small cohort

3.1. Plain (Utterance-level) T-Norm: Results (iii)

- Other causes for limited performance of T-Norm?
  - M. Hébert and D. Boies, (ICASSP’05) analyzed the effect of lexical mismatch, and proposed it as a cause for the poor performance
    - Smoothing mechanism that weighted the effect of T-Norm according to the goodness of the cohort to model the utterance to verify
- Could we reduce the effect of the lexical mismatch in other ways?
  - Reducing the lexical content of the test speech used to produce a speaker verification score to a single phoneme or sub-phoneme
  - And then T-Normalizing these scores and combining them
- Basic idea of Phoneme and Sub-phoneme-level T-Norm
### 3.2. Phoneme-level T-Norm: Procedure

- **Compute phoneme-based verification scores** for the same test utterance, the speaker model and a cohort of impostor models
  - Compute a verification score for each non-silence phoneme $i$, $sc_p(O, \lambda_D, i)$
    - Considering only acoustic scores associated to phoneme $i$ in the utterance
  - Reserve a cohort of impostor speakers $\{1, \ldots, M\}$
  - Obtain MLLR speaker-adapted phonetic HMMs for those speakers
  - For each non-silence phoneme $i$, compute verification scores for the same test utterance and those speaker models $sc_p(O, \lambda^1_D, i), \ldots, sc_p(O, \lambda^M_D, i)$
- **Normalize each phoneme-based verification score** using the mean and standard deviation of the corresponding impostor scores obtained
- Combine normalized phoneme-based verification scores to form utterance verification score (taking into account phoneme lengths)

\[
sc_D(O, \lambda_D) = \frac{1}{N} \sum_{i=1}^{K} N(i)sc_p(O, \lambda_D, i)
\]

### 3.2. Phoneme-level T-Norm: Results

- **Phoneme-level T-Norm vs. No T-Norm on YOHO**
  - Enrolment with only 6 utterances from 1 session and test with 1 utterance
  - 10 male and 10 female speakers reserved as cohort and not included in results
  - Cohort = 20 speaker models
  - MLLR adaptation
- Phoneme-Level T-Norm is clearly better than No T-Norm
- Also clearly better than Utterance-Level T-Norm
- Can we do it better by using even smaller units?
3.3. Subphoneme-level T-Norm: Procedure & Results

- Using exactly the same idea of phoneme-level T-Norm
  - But using HMM states instead of phonemes
- State-level T-Norm vs. No T-Norm on YOHO
  - Enrolment with only 6 utterances from 1 session and test with 1 utterance
  - 10 male and 10 female speakers reserved as cohort and not included in results
  - Cohort = 20 speaker models
  - MLLR adaptation
- Results are even better than with Phoneme-level T-Norm

4. Summary of Results

<table>
<thead>
<tr>
<th>Type of T-Norm</th>
<th>EER (%) (Rel. Improv. %)</th>
<th>FR@FA=1% (%) (Rel. Improv. %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No T-Norm</td>
<td>4.82% (0.0%)</td>
<td>16.28% (0.0%)</td>
</tr>
<tr>
<td>Utterance-based</td>
<td>5.01% (-3.9%)</td>
<td>17.45% (-7.2%)</td>
</tr>
<tr>
<td>Phoneme-based</td>
<td>3.91% (18.9%)</td>
<td>12.17% (25.2%)</td>
</tr>
<tr>
<td>State-based</td>
<td>3.85% (20.1%)</td>
<td>11.81% (27.5%)</td>
</tr>
</tbody>
</table>

- Utterance-level T-Norm performs worse than doing nothing
- But the newly proposed Phoneme-level and State-level T-Norm provide relative improvements in EER close to 20% and over 25% in FR@FA=1%
5. Discussion (i)

- Phoneme and State-level T-Norm work clearly better than Utterance-level T-Norm in text-dependent SR
  - Utterance-level (or Plain) T-Norm suffers from lexical mismatch

- But this mismatch is not totally avoided by Phoneme or State-level T-Norm
  - It is still possible to have substantial differences in lexical content
  - However, now each phoneme/sub-phoneme in the test utterance produces an independent speaker verification score
    - For which the mismatch is limited to the mismatch in a single phoneme/sub-phoneme in the training material
  - This may reduce the influence of the lexical mismatch on the phoneme/sub-phoneme verification scores
  - Making T-Norm less sensitive to this problem

5. Discussion (ii)

- Other possible reason for the good performance of phoneme and state-level T-Norm
  - Based on ideas from a recent paper [Subramanya et al., ICASSP’07]
    - Subramanya computes speaker verification scores for each phoneme
    - And considers those scores as produced by independent weak speaker recognizers
    - That are combined using boosting to yield improved performance
  - This is (conceptually) similar to our approach
    - We combine phoneme or sub-phoneme verification scores
    - Weighting them according to their means and variances on a cohort

- Different phonemes/sub-phonemes → different discriminating powers
  - T-Norm at the phoneme or sub-phoneme levels could be able to weight them appropriately
6. Conclusions

- Applying T-Norm in text-dep SR the way we do in text-indep SR does not work well
  - This is Plain or Utterance-level T-Norm

- Newly proposed T-Norm schemes working at sub-utterance levels work much better
  - Phoneme-level T-Norm
  - Subphoneme-level T-Norm

- Possible reasons
  - Reduction of the effect of lexical mismatch
  - Better weighting/fusion of the information provided by the different phonemes or subphonemes

Thanks!
Baum-Welch Reestimation (YOHO)

<table>
<thead>
<tr>
<th>number of iterations</th>
<th>Gaussians / State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.6, 6.0, 6.8, 7.3, 7.4</td>
</tr>
<tr>
<td>4</td>
<td>6.4, 7.9, 10.0, 14.4, 16.6</td>
</tr>
</tbody>
</table>

- Phonetic HMMs from 1 to 5 Gaussians/State
- Baum-Welch Reestimation
  - 1 or 4 iterations
- 6 enrollment utterances (1 session)
MLLR Adaptation Results (YOHO)

<table>
<thead>
<tr>
<th>Gaussians / State</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.5</td>
<td>6.0</td>
<td>5.9</td>
<td>5.8</td>
<td>5.6</td>
</tr>
<tr>
<td>2</td>
<td>5.3</td>
<td>4.8</td>
<td>4.7</td>
<td>4.6</td>
<td>4.3</td>
</tr>
<tr>
<td>4</td>
<td>9.1</td>
<td>5.6</td>
<td>4.8</td>
<td>4.5</td>
<td>4.2</td>
</tr>
<tr>
<td>8</td>
<td>9.1</td>
<td>5.4</td>
<td>5.1</td>
<td>4.6</td>
<td>4.2</td>
</tr>
<tr>
<td>16</td>
<td>9.1</td>
<td>5.4</td>
<td>4.9</td>
<td>4.7</td>
<td>4.2</td>
</tr>
<tr>
<td>32</td>
<td>9.1</td>
<td>5.4</td>
<td>4.9</td>
<td>4.7</td>
<td>4.2</td>
</tr>
</tbody>
</table>

- Phonetic HMMs with 5, 10, 20, 40 and 80 Gauss/state
- MLLR Adaptation
  - 1, 2, 4, 8, 16, 32 regression classes
- 6 enrollment utterances (1 session)