Transfer Learning for Multi-Person and Multi-Dialect Spoken Language Interface

Mark Hasegawa-Johnson, David Harwath, Harsh Vardhan Sharma, and Po-Sen Huang

University of Illinois

Urbana Neuroengineering Conference, June 28, 2012
Outline

1. Transfer Learning
2. Component Sharing: Automatic Language Identification
3. Parameter Sharing: Automatic Recognition of Dysarthric Speech
4. Data Sharing: Phone Classification in Lebanese Arabic
5. Conclusions
Outline

1 Transfer Learning

2 Component Sharing: Automatic Language Identification

3 Parameter Sharing: Automatic Recognition of Dysarthric Speech

4 Data Sharing: Phone Classification in Lebanese Arabic

5 Conclusions
Humans routinely transfer relevant knowledge from a known task to an unknown task.

Example: A girl who has studied ballet can learn to figure skate more quickly than other girls.
Fundamentals of Machine Learning

- **What is Random:**
  - **The Test Data:** \((x, y)\) are drawn from unknown distribution \(P(x, y)\), where \(x \in \mathcal{X}\) is observation, \(y \in \mathcal{Y}\) is the unknown correct label.
  - **The Training Data:** \(\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}\) are drawn i.i.d. from \(P(x, y)\).

- **What is Learned:** The hypothesis function \(h(x)\) is selected from a function space \(\mathcal{H} : \mathcal{X} \rightarrow \mathcal{Y}\) with covering number \(N_{\mathcal{H}}\) (covering number \(=\) number of meaningfully distinct hypotheses).

- **How it is Evaluated:** Selecting \(h(x)\), if the true label is \(y\), incurs cost \(\ell(h(x), y) \in [0, 1]\).

The Goal of Machine Learning

Choose \(h(x) \in \mathcal{H}\) in order to minimize \(R(h) = E_P [\ell(h(x), y)]\).
Probably Approximately Correct (PAC) Learning

- **Risk:**
  \[
  R(h) = E_P [\ell(h(x), y)]
  \]

- **Empirical Risk Estimate:**
  \[
  \hat{R}(h, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)
  \]

- **Law of Large Numbers:**
  \[
  \hat{R}(h, \mathcal{D}) \rightarrow_{n \to \infty} R(h)
  \]

- **PAC Bound (Haussler, 1983):** with probability at least \(1 - \delta\),
  \[
  \max_{h \in \mathcal{H}} \left| R(h) - \hat{R}(h, \mathcal{D}) \right| \leq \sqrt{\frac{\ln 2N_{\mathcal{H}} - \ln \delta}{2n}}
  \]
How to Use Regularized Machine Learning

1. **Before you see the data**, choose a continuum of possible hypothesis spaces, $\mathcal{H}(\lambda)$ for some regularization parameter $\lambda$, such that $N_{\mathcal{H}}$ is a monotonically increasing function of $\lambda$.

2. **After seeing the data**, choose $\lambda$ and $h \in \mathcal{H}(\lambda)$ in order to minimize

   $$\hat{R}(h, D) + \sqrt{\frac{\ln 2N_{\mathcal{H}} - \ln \delta}{2n}}$$

How Should You Design the Regularizer?

The way in which $\mathcal{H}$ depends on $\lambda$ can have a big impact on performance. How can you design the regularizing function without looking at the training data?

- **Prior Knowledge** can specify $\mathcal{H}(\lambda)$
- **Semi-Supervised Learning**: use unlabeled data
- **Transfer Learning**: use knowledge from a different task
Types of Transfer Learning

- **Component Sharing:**
  - Some **prerequisite skills** of ballet and ice skating are **identical**
  - Example: balance, strength, precise control

- **Parameter Sharing:**
  - **Parameter settings** of ballet and ice skating are **similar, but not identical**; when learning to ice skate, you can initialize with the ballet parameter settings
  - Example: foot position, leg position, pirouette, arabesque

- **Data Sharing:**
  - Moderate **cross-training** can improve performance even when you’re already skilled by **increasing the variability of your training regimen** (improves your robustness to minor perturbations)
  - Example: ballet can improve posture of an ice skater by forcing her to think about posture in a different environment
Transfer Learning by Component Sharing

- $h(x) = h_1(h_2(x))$, such that $N_H = N_1 \times N_2$
- $h_2(x)$ is a **skill applicable to many tasks**, so it is trained using $m \gg n$ training observations
- $h_1$ is trained exclusively **in-task**, using only $n$ training observations—but $h_1$ is a relatively low-complexity function, so $N_1 \ll N_2$
- The PAC bound is reduced from $\sqrt{\frac{\ln N_H + \ln \frac{1}{\delta}}{2n}}$ to
  $$\sqrt{\frac{\ln 2N_1 + \ln \frac{1}{\delta}}{2n}} + \sqrt{\frac{\ln 2N_2 + \ln \frac{1}{\delta}}{2m}}$$
Transfer Learning by Parameter Sharing

- The skill of ice skating is $h(x)$; the skill of ballet is $h_0(x)$. We believe that ice skating is similar to ballet, i.e., that $D(h, h_0)$ is small, for some distance metric $D$.
- Define $\mathcal{H}(\lambda)$ to be the set of all hypotheses, $h(x)$, such that $D(h, h_0) \leq \lambda$.
- Choose $\lambda$ and $h \in \mathcal{H}(\lambda)$ to minimize

$$\hat{R}(h, D) + \sqrt{\frac{\ln 2N_{\mathcal{H}} - \ln \delta}{2n}}$$
Transfer Learning by Data Sharing

- Most probability distributions $P(x, y)$ have infrequent outliers (e.g., Gaussian). Therefore, the training dataset $\mathcal{D}$ has few outliers; if an outlier occurs during testing, then the learned function $h(x)$ will perform very badly. *Example*: the ice skater trips.

- **Moderate outliers** can be added to the training dataset by adding a moderate number of tokens ($k < n$, e.g., $k \approx 0.05n$) from a related task.

- Thus instead of minimizing $R(h) = E_P[\ell(h(x), y)]$, we minimize a robust-risk $R_{\text{robust}}(h) = E_{nP+kQ}^{n+k} [\ell(h(x), y)]$, where $P$ is the in-task distribution (ice skating), and $Q$ is the related-task distribution (ballet).
Examples of Transfer Learning

- **Component Sharing:** Phonological distinctive feature detectors, trained in one language, improve language ID across multiple languages.

- **Parameter Sharing:** Constrained convex search finds optimum bias parameters for speaker adaptation; improve ASR for talkers with severe disability.

- **Data Sharing:** Data from standard Arabic improve the robustness of ASR in Lebanese Arabic.
Outline

1. Transfer Learning

2. Component Sharing: Automatic Language Identification

3. Parameter Sharing: Automatic Recognition of Dysarthric Speech

4. Data Sharing: Phone Classification in Lebanese Arabic

5. Conclusions
Automatic Language Identification: UBM-MAP Framework

UBM-MAP Classification Criterion

Declare that language $L_j$ is present in utterance $X_j$ if

$$\sum_{x_t \in X_j} \ln \left( \frac{\sum_{k=1}^{K} c_k^{(L_j)} \mathcal{N}(x_t; \mu_k^{(L_j)}, \Sigma_k^{(L_j)})}{\sum_{k=1}^{K} c_k^{(UBM)} \mathcal{N}(x_t; \mu_k^{(UBM)}, \Sigma_k^{(UBM)})} \right) > \text{threshold}$$

UBM-MAP Training (Reynolds and Rose, 1995)

Universal background model (UBM) is a $K = 1024$ Gaussian mixture model, trained using data from 15 languages. Language-dependent models are adapted as

$$\mu_k^{(L_j)} = \frac{\tau \mu_k^{(UBM)} + \sum_t \gamma_k(t)x_t}{\tau + \sum_t \gamma_k(t)}$$

Feature vector $x_t$ contains shifted-delta-cepstra (SDC).
Improved Language ID: Distinctive Features

(Harwath and Hasegawa-Johnson, Speech Prosody 2010)
## Distinctive Feature Detector SVMs: Trained Using TIMIT

<table>
<thead>
<tr>
<th>Distinctive Feature</th>
<th>Phoneme Labels (+)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alveolar Fricative</td>
<td>/s/, /z/</td>
<td>82.56%</td>
</tr>
<tr>
<td>Alveolar Nasal</td>
<td>/n/</td>
<td>80.12%</td>
</tr>
<tr>
<td>Alveolar Stop</td>
<td>/d/, /t/</td>
<td>79.68%</td>
</tr>
<tr>
<td>Interdental Fricative</td>
<td>/th/, /dh/</td>
<td>77.70%</td>
</tr>
<tr>
<td>Labial Nasal</td>
<td>/m/</td>
<td>79.50%</td>
</tr>
<tr>
<td>Labial Stop</td>
<td>/b/, /p/</td>
<td>75.62%</td>
</tr>
<tr>
<td>Labiodental Fricative</td>
<td>/f/, /v/</td>
<td>76.38%</td>
</tr>
<tr>
<td>Postalveolar</td>
<td>/jh/, /ch/, /sh/, /zh/</td>
<td>81.76%</td>
</tr>
<tr>
<td>Retroflex</td>
<td>/r/</td>
<td>79.50%</td>
</tr>
<tr>
<td>Velar Stop</td>
<td>/g/, /k/</td>
<td>79.33%</td>
</tr>
<tr>
<td>Sonorant</td>
<td>/iy/, /ih/, /eh/, /ey/, /ae/, /aa/, /aw/, /ay/, /ah/, /ao/, /oy/, /ow/, /uh/, /uw/, /ux/, /er/, /ax/, /ix/, /axr/, /ax-h/, /l/, /r/, /w/, /y/, /hh/, /hv/, /el/, /m/, /n/, /ng/, /em/, /en/, /eng/, /nx/</td>
<td>89.32%</td>
</tr>
</tbody>
</table>
Use of Distinctive Features in Language Detection: Results

Equal Error Rates: 25% GMM, 19% DFSVM-GMM.
Dataset: NIST Callfriend.
Conclusions: Automatic Language Identification

- By augmenting the SDC vector with phonological distinctive features, we reduce EER of automatic language ID to 19% (competitive with the state of the art)
- Distinctive feature labels can only be trained using data with labeled phone boundaries—only available in a handful of languages
- Solution: Train distinctive feature detectors on languages with available data, *use* them for all languages

Future Work

- Apply the method to automatic speech recognition in multiple languages
Outline

1. Transfer Learning
2. Component Sharing: Automatic Language Identification
3. Parameter Sharing: Automatic Recognition of Dysarthric Speech
4. Data Sharing: Phone Classification in Lebanese Arabic
5. Conclusions
Dysarthric Speech

- Motor disorders like Cerebral Palsy can prevent use of a keyboard and mouse.

- *Dysarthria* (articulatory motor disorder) prevents use of standard speech recognition.

- Speaker-dependent ASR can be faster than a keyboard.

- Training a speaker-dependent ASR is difficult, especially for speakers with perpetual muscle spasticity.
Database: UA-Speech (Kim, Hasegawa-Johnson et al., 2008)

**Database**

- 17 talkers with dysarthria
- 3 blocks of isolated words, 255 words/block
- http://isle.illinois.edu/sst/data/UASpeech

**Data Example**
Speaker-Dependent Training

Given posterior probabilities
\[ \gamma_k(t) = p(\text{Gaussian } k|\text{Observation } x_t), \]
the Gaussian means are updated as
\[
\mu_k^{(SD)} = \frac{\sum_t \gamma_k(t)x_t}{\sum_t \gamma_k(t)}
\]

Speaker Adaptation

Given a speaker-independent mean vector \( \mu_k^{(SI)} \) and regularization constant \( \tau \),
\[
\mu_k^{(SA)} = \frac{\tau \mu_k^{(SI)} + \sum_t \gamma_k(t)x_t}{\tau + \sum_t \gamma_k(t)}
\]
Recognition Accuracy: SD vs. SA ASR
(Sharma & Hasegawa-Johnson, SLPAT 2010)

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Intelligibility (%)</th>
<th>WRA:SD (%)</th>
<th>WRA:SA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M09</td>
<td>86</td>
<td>52</td>
<td>64</td>
</tr>
<tr>
<td>M05</td>
<td>58</td>
<td>36</td>
<td>38</td>
</tr>
<tr>
<td>M06</td>
<td>39</td>
<td>34</td>
<td>41</td>
</tr>
<tr>
<td>F02</td>
<td>29</td>
<td>35</td>
<td>30</td>
</tr>
<tr>
<td>M07</td>
<td>28</td>
<td>44</td>
<td>39</td>
</tr>
<tr>
<td>F03</td>
<td>6</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>M04</td>
<td>2</td>
<td>2.8</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Training data: blocks 1 and 3 (510 word tokens, 355 word types).
Test data: block 2 (test vocabulary=test tokens=255 words).
Improved Adaptive Training: Background Interpolation

\[
\mu_k^{(BI)} = \frac{\tau (\delta \mu_k^{(SI)} + (1 - \delta) \mu_k^{(SDB)}) + \sum_t \gamma_k(t) x_t}{\tau + \sum_t \gamma_k(t)}
\]

Speaker-Dependent Background Model (SDB)

**SDB** is a three-state HMM representing *all phones* (data from all phones pooled) as produced by the target speaker (inspiration: speaker recognition models). Able to capture phone-independent speaker characteristics including speaking rate, voice quality.
## Recognition Accuracy: BI-Map

(Sharma & Hasegawa-Johnson, in review)

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Intel. (%)</th>
<th>WRA:SD (%)</th>
<th>WRA:SA (%)</th>
<th>WRA:BI-Map (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M09</td>
<td>86</td>
<td>52</td>
<td>64</td>
<td>72</td>
</tr>
<tr>
<td>M05</td>
<td>58</td>
<td>36</td>
<td>38</td>
<td>39</td>
</tr>
<tr>
<td>M06</td>
<td>39</td>
<td>34</td>
<td>37</td>
<td>41</td>
</tr>
<tr>
<td>F02</td>
<td>29</td>
<td>35</td>
<td>29</td>
<td>33</td>
</tr>
<tr>
<td>M07</td>
<td>28</td>
<td>44</td>
<td>39</td>
<td>43</td>
</tr>
<tr>
<td>F03</td>
<td>6</td>
<td>13</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>M04</td>
<td>2</td>
<td>2.8</td>
<td>3.0</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Training data: blocks 1 and 3 (510 word tokens, 355 word types). Test data: block 2 (test vocabulary=test tokens=255 words).
Conclusions: Dysarthric Speech Recognition

- Speaker adaptation is better than speaker-dependent training for most speakers (given the amount of training data used in this study).

- Background interpolation provides extra flexibility to the MAP algorithm, which can be used to generate extra accuracy.

- Accuracies with 255-word vocabulary are still too low for practical use. Green et al. (Sheffield U.) are experimenting with interactive interface design methods to set vocabulary size.
Outline

1. Transfer Learning
2. Component Sharing: Automatic Language Identification
3. Parameter Sharing: Automatic Recognition of Dysarthric Speech
4. Data Sharing: Phone Classification in Lebanese Arabic
5. Conclusions
Cross-Dialect Transfer Learning

- The Arabic language can be viewed as a family of related languages, with limited vocabulary overlap between dialects, but with a high percent overlap among the phoneme inventories.

- The idea of cross-dialectal transfer learning is to bridge the knowledge from one dialect to the other, assuming that different dialects still have knowledge in common.
Cross-Dialectal Training Objective

- Suppose we are given a set of points $X_{D_p} = \{x_1, \ldots, x_n\}$ with labels $Y_{D_p} = \{y_1, \ldots, y_n\}$ and another set of points $X_{D_q} = \{x_{n+1}, \ldots, x_{n+k}\}$ with labels $Y_{D_q} = \{y_{n+1}, \ldots, y_{n+k}\}$, where $D_p$ and $D_q$ are datasets from two different dialects.

- In our case, $x_i \in \mathbb{R}^d$ is the $d$-dimensional spectral feature vector, and $y_i \in \{1, \ldots, C\}$ is the phone class label, assuming that there are $C$ phone classes.

- The classification rule $\mathbb{R}^d \rightarrow \{1, \ldots, C\}$ is based on Bayes’ rule,

$$\hat{y} = \arg \max p(x|y)p(y),$$

where $p(y)$ is the phone class prior estimated from the training set, and the conditional distribution $p(x|y), y \in \{1, \ldots, C\}$ is modeled using a Gaussian mixture model (GMM).
Cross-Dialectal Training Objective
(Huang and Hasegawa-Johnson, CITALA 2012)

- GMM training objective:

\[
J = \mathcal{L}(\theta_c | X_{Dp,c}) + \alpha \sum_{j=1}^{C} P(Y_{Dp,c} = Y_{Dq,j}) \mathcal{L}(\theta_c | X_{Dq,j}) \quad (1)
\]

\[
= \sum_{i=1}^{n} \ln p(x_i | \theta_c) + \alpha \sum_{i=n+1}^{n+k} p(y_i = c) \ln p(x_i | \theta_c) \quad (2)
\]

where \( \mathcal{L}(\theta | X_{D,j}) \) is the likelihood of data \( X_{D,j} \) corresponding to phone \( j \) in dialect \( D \) under the assumed model \( \theta \).

- For simplicity we use \( \alpha = 1 \) and \( p(y_i = c) \) equal to a delta function, i.e., \( p(y_i = c) = 1 \) if \( y_i \) and \( c \) share an IPA symbol, \( p(y_i = c) = 0 \) otherwise.

- Maximum likelihood model parameters are estimated using the EM algorithm.
MSA Corpus: West Point MSA

- West Point Modern Standard Arabic corpus
- There are 8,516 speech files, totaling 11.42 hours or 1.7 GB of speech data. Each speech file is recorded by one subject reciting one prompt from one of four prompt scripts.
- Approximately 7,200 files are from native speakers and 1,200 files are from non-native speakers.
- There are totally 1,131 distinct Arabic words. All scripts were written with MSA and were diacriticized.
BBN/AUB DARPA Babylon Levantine Arabic Speech corpus

Levantine Arabic is an Arabic dialect of people in Lebanon, Jordan, Palestine, and Syria.

It is a spoken language instead of a written language. About 1/3 of word tokens are not in a typical MSA lexicon; many words shared by MSA are pronounced differently.

The BBN/AUB dataset consists of 164 speakers, 101 males and 63 females. It is a set of spontaneous speech sentences, recorded in Levantine colloquial Arabic.

The duration of recorded speech is 45 hours distributed among 79,500 audio clips.
Experiment Settings

- We first use forced alignment to generate phone boundary information.
- Then, for each phone token, 12-dim PLP features are calculated with 25 ms Hamming window and 10 ms frame rate. Three average PLP vectors are computed, averaged over each third of the phone token, and concatenated to form a segmental feature vector.
- For classification task, we omit the glottal stop phone in both corpora. There are 36 phone classes in West Point Modern Standard Arabic Speech corpus and 38 phone classes in BBN/AUB DARPA Babylon Levantine Arabic Speech corpus.
- For Levantine corpus, we randomly choose 60% of the data as training set, 10% of the data as development set, and the remaining 30% of the data as test set. We use the whole set of MSA data for transfer learning.
Results: Transfer MSA data to Levantine Arabic

Phone classification accuracy, system trained using a fraction of the MSA data and a fraction of the Levantine data. **Abscissa:** fraction of the Levantine corpus used for training (2.7 to 27 hours). **Parameter:** fraction of the MSA corpus used for training (0 to 1.1 hours). **Ordinate:** phone accuracy, Levantine test data.
Results: Transfer MSA data to Levantine Arabic

Same data as previous slide, but replotted with **Abscissa** = ratio between the length of Levantine training data and MSA data. All systems peak at a data ratio of 20:1 (Levantine:MSA).
Discussion/Conclusions

- Low classification accuracy: Force-aligned phone boundaries are not precise. Hence, the feature vector representing a phone token might also include features from its adjacent phones.
- The classification accuracy depends on the ratio between the length of Levantine training data and MSA data.
  - When proper ratio of MSA data is transferred to Levantine data (MSA data 1/20 of Levantine data), we enhance the phonetic coverage of GMMs and achieve higher accuracies.
  - Transferring too much MSA data apparently results in mis-matched phone acoustic models.
- Future Work: Test other transfer learning methods including parameter-sharing methods (MAP-adaptation-like), metric-sharing methods (distinctive feature/landmark-based systems).
Outline

1. Transfer Learning
2. Component Sharing: Automatic Language Identification
3. Parameter Sharing: Automatic Recognition of Dysarthric Speech
4. Data Sharing: Phone Classification in Lebanese Arabic
5. Conclusions
Training in one task can improve one’s learning rate in a second task. There are at least three ways this can be accomplished:

1. **Component Sharing**, e.g., distinctive feature detectors useful in English are not sufficient to represent all sounds of all languages, but are useful to represent some of them.

2. **Parameter Sharing**, e.g., parameters of a talker-adaptive acoustic model may be initialized by interpolating between the talker-independent HMM and the talker-dependent GMM.

3. **Data Sharing**, e.g., phone classification data in Standard Arabic can improve robustness of a Levantine Arabic phone classifier.