When mismatched training data outperform matched data

Emmanuel Vincent

To cite this version:

Emmanuel Vincent. When mismatched training data outperform matched data. Systematic approaches to deep learning methods for audio, Sep 2017, Vienna, Austria. hal-01588876

HAL Id: hal-01588876
https://inria.hal.science/hal-01588876
Submitted on 17 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
WHEN MISMATCHED TRAINING DATA OUTPERFORM MATCHED DATA

Emmanuel Vincent
Inria Nancy – Grand Est, France
Noise-robust speech recognition

Sitting in a cafe (CAF)

Standing at a street junction (STR)

Traveling on a bus (BUS)

In a pedestrian area (PED)
Some applications

Tasks: speech, speaker, or language recognition, paralinguistics. . .
Challenges and approach

Challenges:
- reverberation
- multiple, nonstationary noise sources
- overlapping speech
- moving sources and/or microphones.

General approach:
- single- or multichannel speech enhancement/separation
- combined with better features and acoustic model.
DEEP LEARNING BASED SPEECH ENHANCEMENT
Single-channel enhancement/separation

Spectral filtering achieved via time-frequency masking.

Speech source

Speech + noise mixture

Spectral filter

Filtered signal
Using DNNs for single-channel separation

Training data: simulated mixtures of speech and noise
Test data: time-frequency mask computed as

\[
\frac{v_{\text{speech}}(f, n)}{v_{\text{speech}}(f, n) + v_{\text{noise}}(f, n)}
\]
Results

Multichannel enhancement/separation

Combination of spatial and spectral filtering.

Spatial filter (anechoic)

Spatial filter (reverberant)
Using DNNs for multichannel separation in an EM fashion

\[ x(f, n) = \sum_j c_j(f, n) \]

\[ c_j(f, n) \sim \mathcal{N}(0, v_j(f, n) R_j(f)) \]

Noisy speech

Spatial image of j-th source

Power spectrum

Spatial covariance matrix

**Parameter updates:**

- Update spatial covariance matrix:
  \[ R_j(f) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{v_j(f, n)} \hat{R}_{c_j}(f, n) \]

- Compute unconstrained spectrogram:
  \[ z_j(f, n) = \frac{1}{I} \text{tr} \left( R_j(f)^{-1} \hat{R}_{c_j}(f, n) \right) \]

- Update spectrogram given DNN:
  \[ v_j(f, n) \leftarrow \text{DNN} (z_j(f, n)) \]
Performance across iterations

Spatial filtering translates into improved DNN inputs, which result in improved outputs.

SDR ($\Rightarrow$ distortion)  SIR ($\Rightarrow$ interference)  SAR ($\Rightarrow$ artifact)
## Results

Evaluation in terms of word error rate (WER) with 6 mics.

<table>
<thead>
<tr>
<th>Method</th>
<th>WER Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td></td>
</tr>
<tr>
<td>Single-channel DNN</td>
<td>no WER reduc.</td>
</tr>
<tr>
<td>Delay-and-sum</td>
<td>21% rel. WER reduc.</td>
</tr>
<tr>
<td>DNN post-filter</td>
<td>20% rel. WER reduc.</td>
</tr>
<tr>
<td>Multichannel DNN</td>
<td>39% rel. WER reduc.</td>
</tr>
</tbody>
</table>

DEEP LEARNING BASED SPEECH RECOGNITION
DNN-based acoustic modeling

DNNs are also the state-of-the-art for acoustic modeling of speech

- input features: MFCCs, logmel, waveform
- outputs: phonetic classes, characters
Full speech recognition system

Robust automatic speech recognition (ASR) systems typically involve:

- multichannel speech enhancement
- robust features
  - auditory-inspired
  - feature adaptation (fMLLR)
- robust acoustic modeling
  - CNN, BLSTM
  - training data augmentation (more SNRs, enhanced data)
  - model adaptation
- system combination
CHiME-3/CHiME-4 benchmark

CHiME-3 dataset: WSJ0 utterances in four noise environments.

Original data (similar noises types and SNRs across all datasets):
- training: 7138 simulated utterances
- development: 1640 real + 1640 simulated utterances
- test: 1320 real utterances.

Baseline: feedforward DNN acoustic model on fMLLR features + 3-gram language model. No speech enhancement.

WER reduced from 33.4% (baseline) to 2.2% (best). Close to clean speech performance, possibly due to matched hardware and “all-inclusive” training data.

How much do environment or microphone mismatch affect enhancement and ASR performance?
IMPACT OF ENVIRONMENT AND MICROPHONE MISMATCHES
Environment mismatch

Noise characteristics vary within and across environments.
Environment mismatch — Impact on enhancement

WER achieved by multichannel enhancement (with baseline acoustic model)

<table>
<thead>
<tr>
<th>Training (real)</th>
<th>Test (real)</th>
<th>BUS</th>
<th>CAF</th>
<th>PED</th>
<th>STR</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUS</td>
<td>21.03</td>
<td>13.06</td>
<td>17.92</td>
<td>9.28</td>
<td>15.32</td>
<td></td>
</tr>
<tr>
<td>CAF</td>
<td>31.48</td>
<td>13.15</td>
<td>16.95</td>
<td>8.78</td>
<td>17.59</td>
<td></td>
</tr>
<tr>
<td>PED</td>
<td>27.89</td>
<td>12.20</td>
<td>17.04</td>
<td>8.93</td>
<td>16.51</td>
<td></td>
</tr>
<tr>
<td>STR</td>
<td>24.30</td>
<td>11.80</td>
<td>16.42</td>
<td>8.48</td>
<td>15.25</td>
<td></td>
</tr>
<tr>
<td>1/4 of all</td>
<td>20.83</td>
<td>11.65</td>
<td>15.94</td>
<td>8.72</td>
<td>14.28</td>
<td></td>
</tr>
<tr>
<td>all but BUS</td>
<td>22.62</td>
<td>10.72</td>
<td>15.47</td>
<td>7.55</td>
<td>14.09</td>
<td></td>
</tr>
<tr>
<td>all but CAF</td>
<td>18.90</td>
<td>10.59</td>
<td>16.07</td>
<td>7.53</td>
<td>13.27</td>
<td></td>
</tr>
<tr>
<td>all but PED</td>
<td>18.56</td>
<td>10.76</td>
<td>14.93</td>
<td>8.09</td>
<td>13.08</td>
<td></td>
</tr>
<tr>
<td>all but STR</td>
<td>18.19</td>
<td>10.03</td>
<td>15.08</td>
<td>7.94</td>
<td>12.81</td>
<td></td>
</tr>
<tr>
<td>3/4 of all</td>
<td>18.84</td>
<td>10.98</td>
<td>15.41</td>
<td>7.79</td>
<td>13.26</td>
<td></td>
</tr>
</tbody>
</table>

1 training environment:
- Multicondition: 14.28%
- Matched: 14.93%
- Mismatched: 16.58%

3 training environments:
- Multicondition: 13.26%
- Mismatched: 14.02%

⇒ multicondition training preferable to matched training
⇒ on average, performs well on environments not seen in training
Environment mismatch — Impact on ASR

WER achieved by acoustic model (no enhancement, RNN language model)

<table>
<thead>
<tr>
<th>Training (real + sim)</th>
<th>Test (real)</th>
<th>BUS</th>
<th>CAF</th>
<th>PED</th>
<th>STR</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUS</td>
<td></td>
<td>45.56</td>
<td>33.34</td>
<td>26.53</td>
<td>17.71</td>
<td>30.78</td>
</tr>
<tr>
<td>CAF</td>
<td></td>
<td>44.33</td>
<td>23.22</td>
<td>18.78</td>
<td>16.88</td>
<td>25.80</td>
</tr>
<tr>
<td>PED</td>
<td></td>
<td>43.86</td>
<td>23.53</td>
<td>17.53</td>
<td>17.37</td>
<td>25.57</td>
</tr>
<tr>
<td>STR</td>
<td></td>
<td>40.31</td>
<td>28.63</td>
<td>22.27</td>
<td>16.14</td>
<td>26.83</td>
</tr>
<tr>
<td>1/4 of all</td>
<td></td>
<td>40.47</td>
<td>23.52</td>
<td>17.90</td>
<td>15.47</td>
<td>24.34</td>
</tr>
<tr>
<td>all but BUS</td>
<td></td>
<td>35.50</td>
<td>18.34</td>
<td>13.66</td>
<td>12.29</td>
<td>19.94</td>
</tr>
<tr>
<td>all but CAF</td>
<td></td>
<td>32.87</td>
<td>20.92</td>
<td>15.45</td>
<td>12.25</td>
<td>20.37</td>
</tr>
<tr>
<td>all but PED</td>
<td></td>
<td>32.62</td>
<td>20.64</td>
<td>15.66</td>
<td>12.33</td>
<td>20.31</td>
</tr>
<tr>
<td>all but STR</td>
<td></td>
<td>33.11</td>
<td>18.88</td>
<td>15.06</td>
<td>12.94</td>
<td>20.00</td>
</tr>
<tr>
<td>3/4 of all</td>
<td></td>
<td>32.75</td>
<td>19.41</td>
<td>13.45</td>
<td>12.40</td>
<td>19.50</td>
</tr>
</tbody>
</table>

1 training environment:
- Multicondition: 24.34%
- Matched: 25.61%
- Mismatched: 27.80%

3 training environments:
- Multicondition: 19.50%
- Mismatched: 21.26%

⇒ similarly small impact on ASR as on enhancement

⇒ again, multicondition training performs best on average
Microphone response mismatch

Relative responses obtained by averaging over 1 min segments.
### Microphone response mismatch — Impact on ASR

WER achieved by acoustic model (no enhancement, RNN language model)

<table>
<thead>
<tr>
<th>Training (real + sim)</th>
<th>Test (real)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ch1</td>
<td>ch1</td>
<td>ch2</td>
<td>ch3</td>
<td>ch4</td>
<td>ch5</td>
<td>ch6</td>
</tr>
<tr>
<td>ch1</td>
<td>31.65</td>
<td>83.01</td>
<td>35.20</td>
<td>30.62</td>
<td>27.07</td>
<td>30.76</td>
</tr>
<tr>
<td>ch2</td>
<td>32.43</td>
<td>71.36</td>
<td>35.85</td>
<td>32.44</td>
<td>29.12</td>
<td>31.77</td>
</tr>
<tr>
<td>ch3</td>
<td>30.86</td>
<td>83.04</td>
<td>34.94</td>
<td>30.20</td>
<td>26.26</td>
<td>30.91</td>
</tr>
<tr>
<td>ch4</td>
<td>31.65</td>
<td>82.74</td>
<td>35.64</td>
<td>28.74</td>
<td>25.13</td>
<td>29.27</td>
</tr>
<tr>
<td>ch5</td>
<td>33.60</td>
<td>84.05</td>
<td>36.70</td>
<td>30.25</td>
<td>25.84</td>
<td>30.73</td>
</tr>
<tr>
<td>ch6</td>
<td>31.33</td>
<td>81.65</td>
<td>35.18</td>
<td>28.25</td>
<td>24.72</td>
<td>28.04</td>
</tr>
</tbody>
</table>

Matched: 36.76%
Mismatched: 39.71%

⇒ matched training performs best on average
⇒ performs well on other mics, except backward mic (ch 2)

Is multicondition training the best one can do?
Environment mismatch does not improve performance

WER achieved by acoustic model (no enhancement, RNN language model)

<table>
<thead>
<tr>
<th>Training (real + sim)</th>
<th>BUS</th>
<th>CAF</th>
<th>PED</th>
<th>STR</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUS</td>
<td>45.56</td>
<td>33.34</td>
<td>26.53</td>
<td>17.71</td>
<td>30.78</td>
</tr>
<tr>
<td>CAF</td>
<td>44.33</td>
<td>23.22</td>
<td>18.78</td>
<td>16.88</td>
<td>25.80</td>
</tr>
<tr>
<td>PED</td>
<td>43.86</td>
<td>23.53</td>
<td>17.53</td>
<td>17.37</td>
<td>25.57</td>
</tr>
<tr>
<td>STR</td>
<td>40.31</td>
<td>28.63</td>
<td>22.27</td>
<td>16.14</td>
<td>26.83</td>
</tr>
<tr>
<td>1/4 of all</td>
<td>40.47</td>
<td>23.52</td>
<td>17.90</td>
<td>15.47</td>
<td>24.34</td>
</tr>
</tbody>
</table>

1 training environment:
- Multicondition: 24.34%
- Matched: 25.61%
- Mismatched: 27.80%

3 training environments:
- Multicondition: 19.50%
- Mismatched: 21.26%

Better than multicondition (not significant)

⇒ excluding certain noises does not significantly improve performance
Microphone mismatch does not improve performance

WER achieved by acoustic model (no enhancement, RNN language model)

<table>
<thead>
<tr>
<th>Training (real + sim)</th>
<th>Test (real)</th>
<th>Matched: 36.76%</th>
<th>Mismatched: 39.71%</th>
<th>Better than matched (not significant)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ch1</td>
<td>ch2</td>
<td>ch3</td>
<td>ch4</td>
</tr>
<tr>
<td>ch1</td>
<td>31.65</td>
<td>83.01</td>
<td>35.20</td>
<td>30.62</td>
</tr>
<tr>
<td>ch2</td>
<td>32.43</td>
<td>71.36</td>
<td>35.85</td>
<td>32.44</td>
</tr>
<tr>
<td>ch3</td>
<td>30.86</td>
<td>83.04</td>
<td>34.94</td>
<td>30.20</td>
</tr>
<tr>
<td>ch4</td>
<td>31.65</td>
<td>82.74</td>
<td>35.64</td>
<td>28.74</td>
</tr>
<tr>
<td>ch5</td>
<td>33.60</td>
<td>84.05</td>
<td>37.70</td>
<td>30.25</td>
</tr>
<tr>
<td>ch6</td>
<td>31.33</td>
<td>81.65</td>
<td>35.18</td>
<td>28.25</td>
</tr>
</tbody>
</table>

⇒ training on another microphone does not significantly improve performance either
Data augmentation greatly improves performance

Simulated training data augmented 7 fold by changing the SNR by 
$-15, -10, -5, +5, +10,$ and $+15$ dB relative wrt original.

Speech and noise signals unchanged.

<table>
<thead>
<tr>
<th>Training dataset</th>
<th>Development dataset</th>
<th>WER on test dataset (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BUS</td>
</tr>
<tr>
<td>Original</td>
<td>Simu+real</td>
<td>50.08</td>
</tr>
<tr>
<td>Augment</td>
<td>Simu+real</td>
<td>36.85</td>
</tr>
<tr>
<td>Augment</td>
<td>Real</td>
<td>37.37</td>
</tr>
</tbody>
</table>

⇒ data augmentation improves WER by 8 or 14% relative (highly significant) despite increased SNR mismatch
Other practical evidence

Significant WER improvement also reported in the ASR literature when enhancing the test data, but not the training data.


See also more general proof in the machine learning literature.

Optimizing the training set

What is the optimal training set given the task, the classifier, and the average test conditions (no adaptation)?

So far, tuned by trial and error.
Importance weighting

Optimization framework: importance weighting

- generate a large training set by applying all possibly relevant data augmentation techniques and parameters
- then weight every sample according to its “usefulness”.

Conventional transfer learning: weight each training sample $x$ by $p_{test}(x)/p_{train}(x)$ so that the training and test data are matched.

According to the above evidence, this is suboptimal $\Rightarrow$ discriminative approach required.

Idea: weight every training sample so as to minimize error on a development set.
DISCRIMINATIVE DATA AUGMENTATION
Classically, a DNN is trained to estimate the posterior $p_\theta(y|x)$ over labels (senones) $y$ given inputs $x$.

The parameters $\theta$ are obtained by minimizing the average loss $\mathcal{L}$ on the training set:

$$\hat{\theta} = \arg \min_{\theta} E_{p_{\text{train}}(x)p(y|x)}[\mathcal{L}(p_\theta(y|x), y)].$$

e.g., using cross-entropy:

$$E[\mathcal{L}(p_\theta(y|x), y)] = -\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(y_i|x_i).$$
Proposed discriminative importance weighting objective

Define a weighted loss function with importance weights $\omega_i \geq 0$, e.g., using cross-entropy:

$$E [\mathcal{L}_\omega(p_\theta(y|x), y, \omega)] = - \frac{\sum_{i=1}^{N} \omega_i \log p_\theta(y_i|x_i)}{\sum_{i=1}^{N} \omega_i}.$$  

Solve the following nested optimization problem:

- find the DNN parameters $\hat{\theta}$ that minimize the weighted loss on the training set
- find the data weights $\hat{\omega}$ for which this DNN yields minimum average error rate $E$ on the development set

$$\hat{\theta} = \arg \min_\theta E_{p_{\text{train}}(x)p(y|x)} \left[ \mathcal{L}_\omega(p_\theta(y|x), y, \omega) \right]$$

$$\hat{\omega} = \arg \min_\omega E_{p_{\text{dev}}(x)p(y|x)} \left[ E(p_{\hat{\theta}}(y|x), y) \right].$$
Algorithm

Alternating optimization algorithm:

- update the parameters $\theta$ via one epoch of stochastic gradient descent (SGD) on the full set
  - compute gradient for each $x_i$ by usual backpropagation
  - multiply it by $\omega_i / \sum_{i=1}^{N} \omega_i$ before summing over time

- update the weights $\omega$ using one step of gradient descent
  - for each weight $\omega_i$, update the DNN via one step of SGD on a single sample $x_i$
  - compute resulting difference $\Delta e_i$ in the classification error
  - update the weights as $\omega_i = \omega_i - \lambda \Delta e_i$ with a learning rate $\lambda$

In practice, weights are tied across subsets of data.
Experimental setup

Training set augmented by changing the SNR by $-15$, $-10$, $-5$, $+5$, $+10$, and $+15$ dB relative for the same utterances and noises.

Weights tied across all utterances generated using the same relative SNR w.r.t. the original training set $\Rightarrow 7$ weights in total.
Results

Table 1: ASR WER comparison using acoustic model trained on different train and development datasets with and without applying the weighting algorithm. Composite dataset refers to the data with \{-15, -10, -5, 0, +5, +10, +15\} dB. Testing was done using the real part of the CHiME-3 test dataset.

<table>
<thead>
<tr>
<th>Dataset Type</th>
<th>Weighting Type</th>
<th>Dataset Type</th>
<th># Train utterances</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>No</td>
<td>Simu + Real</td>
<td>7138</td>
<td>50.08</td>
</tr>
<tr>
<td>Composite</td>
<td>No</td>
<td>Simu + Real</td>
<td>49966</td>
<td>36.85</td>
</tr>
<tr>
<td>Composite</td>
<td>Yes</td>
<td>Simu + Real</td>
<td>49966</td>
<td>34.23</td>
</tr>
<tr>
<td>Composite</td>
<td>No</td>
<td>Real</td>
<td>49966</td>
<td>37.37</td>
</tr>
<tr>
<td>Composite</td>
<td>Yes</td>
<td>Real</td>
<td>49966</td>
<td>32.60</td>
</tr>
</tbody>
</table>

Fig. 1: FER on the development set achieved by weighted vs. unweighted training using simulated+real development data. The FER equals 0.999 for both systems at iteration 1.

4.3. Learned data weights

Fig. 2 shows the evolution of the data weights per iteration. The weights for the -10, -5 and -15 dB datasets drop significantly below their initial value, while the weight for the +5 dB dataset increases significantly above its initial value. After 12 iterations, the distribution of the weighted training data is very different from that of the original training data (and from the development and test data) and it focuses on higher SNRs. This further supports the claim that mismatched training data can outperform matched data and that algorithms seeking to select matched training data [15–20] can only achieve limited success. Our algorithm was able to find a “suitably mismatched” distribution of SNRs, a result that would arguably have been hard or impossible to achieve by simple trial and error. Yet, its computational cost is only twice that of training on the unweighted composite dataset.

5. CONCLUSION

In this work we proposed an algorithm to optimize the training set for a DNN acoustic model given a development set that is representative of the test conditions. Our algorithm learns importance weights for disjoint subsets of data in the training set. The learned weights represent the importance of the data samples in each of the subsets towards minimizing the FER in the development set. Experiments were performed on the CHiME-3 dataset by simulating noisy speech with various SNRs. The results show a WER improvement of 15% relative compared to training from the unweighted dataset.

There are two future directions which we believe require further investigation. The first one is to perform a larger-scale experimental evaluation using not only noisy simulated data, but a combination of clean, enhanced, and noisy data and real and simulated data. The second one to investigate if learning weights for random subsets of data (instead of SNR specific subsets in this work) yields any realistic improvements in ASR performance. This is particularly helpful when large amounts of data are available for training with no predefined demarcation information (such as SNR).

6. ACKNOWLEDGMENTS

We acknowledge the support of Bpifrance (FUI voiceHome). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several universities as well as other organizations (see https://www.grid5000.fr).

⇒ resulting SNR distribution ≠ original, hence ≠ test
## Results

<table>
<thead>
<tr>
<th>Training dataset</th>
<th>Development dataset</th>
<th>WER on test dataset (%)</th>
<th>BUS</th>
<th>CAF</th>
<th>PED</th>
<th>STR</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>Simu+real</td>
<td>50.08</td>
<td>27.27</td>
<td>20.37</td>
<td>18.51</td>
<td>29.05</td>
<td></td>
</tr>
<tr>
<td>Augment</td>
<td>Simu+real</td>
<td>36.85</td>
<td>26.84</td>
<td>20.80</td>
<td>15.22</td>
<td>24.92</td>
<td></td>
</tr>
<tr>
<td>Augment+weight</td>
<td>Simu+real</td>
<td>34.23</td>
<td>23.53</td>
<td>19.64</td>
<td>13.82</td>
<td><strong>22.80</strong></td>
<td></td>
</tr>
<tr>
<td>Augment</td>
<td>Real</td>
<td>37.37</td>
<td>30.11</td>
<td>23.22</td>
<td>15.65</td>
<td>26.59</td>
<td></td>
</tr>
<tr>
<td>Augment+weight</td>
<td>Real</td>
<td>32.60</td>
<td>24.15</td>
<td>19.86</td>
<td>14.10</td>
<td><strong>22.68</strong></td>
<td></td>
</tr>
</tbody>
</table>

⇒ discriminative weighting further improves the WER

⇒ but weights not so contrasted…
CONCLUSION
Conclusion

DNNs for speech enhancement or ASR are surprisingly robust to mismatched training data.

SNR-augmented training data improves performance despite SNR mismatch. Not all mismatched training data do.

Discriminative weighting algorithm can potentially improve performance, but efficiency and understanding still to be improved.
References


