SAT-LHUC: Speaker adaptive training for learning hidden unit contributions

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
SAT-LHUC: SPEAKER ADAPTIVE TRAINING FOR LEARNING HIDDEN UNIT CONTRIBUTIONS

Pawel Swietojanski and Steve Renals

The Centre for Speech Technology Research, University of Edinburgh
\{p.swietojanski, s.renals\}@ed.ac.uk

ABSTRACT
This paper extends learning hidden unit contributions (LHUC) unsupervised speaker adaptation with speaker adaptive training (SAT). Contrary to other SAT approaches, the proposed technique does not require speaker-dependent features, the generation of auxiliary generative models to estimate or extract speaker-dependent information, or any changes to the speaker-independent model structure. SAT-LHUC is directly integrated into the objective and jointly learns speaker-independent and speaker-dependent representations. We demonstrate that the SAT-LHUC technique can match feature-space regression transforms for matched narrow-band data and outperform it on wide-band data when the runtime distribution differs significantly from training one. We have obtained 6.5%, 10% and 18.5% relative word error rate reductions compared to speaker-independent models on Switchboard, AMI meetings and TED lectures, respectively. This corresponds to relative gains of 2%, 4% and 6% compared with non-SAT LHUC adaptation. SAT-LHUC was also found to be complementary to SAT with feature-space maximum likelihood linear regression transforms.

Index Terms— SAT, Deep Neural Networks, LHUC

1. INTRODUCTION
Acoustic model (AM) adaptation aims to normalise the mismatch between training and runtime data distributions owing to the acoustic variability among speakers as well as other distortions introduced by the channel or acoustic environment. Speaker adaptive training (SAT) \[1,2\], initially proposed for Gaussian mixture models (GMMs), aims to build a canonical acoustic model that is adjusted to the particular characteristics of speakers using linear transforms (operating in either model space \[3\] or feature space \[4\]) and found by maximising the likelihood of adaptation data under the model. Those techniques are often referred to as Maximum Likelihood Linear Regression (MLLR) transforms and the feature-space variant (fMLLR) has been successfully applied to speaker adaptive training of Deep Neural Network (DNN) acoustic models \[5\] often bringing significant improvements in accuracy \[6,7,8\].

Here we are primarily concerned with direct speaker adaptive training of DNN parameters. Contrary to test-only adaptation approaches \[10,11,12,13,14,15,16,17\], SAT may offer a more tunable canonical DNN model which is able to perform normalisation better than test-only adaptation. At the same time, we are interested in investigating the possibility of SAT training without using auxiliary features (such as i-vectors \[18,19,9,20\]), bottleneck features \[21,22\]) or additional speaker-dependent (SD) parameters that are added to the speaker-independent (SI) model and retuned in a separate SAT phase \[23,24,25,26,27,20,28\].

This paper builds on the recently introduced DNN model-based speaker adaptation technique of learning hidden unit contributions (LHUC) \[15,16\]. In LHUC, an amplitude parameter is introduced for each hidden unit, tied on a per-speaker basis, and estimated in supervised \[15\] or unsupervised \[16\] fashion, the latter using first-pass alignments. This technique has resulted in significant reductions in WER, when tested using the TED talks datasets from the IWSLT evaluation, and was complementary to fMLLR \[16\]. Here, we extend this approach to speaker adaptive training (SAT-LHUC) in which SI and SD LHUC transforms are estimated during training.

2. LHUC AND SPEAKER ADAPTIVE TRAINING
A speaker independent DNN consists of multiple hidden layers, each implementing some non-linear transformations. Each individual hidden unit acts as an adaptive basis function that learns to recognise certain patterns in the previous layer. The learning process for the DNN is driven by a single objective, with the hidden units driven to specialize and become complementary to each other, in order improve the objective. To explain different patterns in the training data the hidden units learn some joint representation of the problem the model was tasked to solve. However, when the model is applied to unseen data, the relative importance of the hidden units may no longer be optimal. LHUC, given adaptation data, rescales the contributions (amplitudes) of the hidden units in the model without actually modifying their feature receptors (Fig. 2). LHUC modifies \( h_j \), the hidden unit output of unit \( j \) in layer \( l \), using a speaker-dependent amplitude function:

\[
h_j = \xi(r_j^{s}) \circ \psi(w_j x + b_j^{s}).
\]

\( r_j^{s} \in \mathbb{R} \) is an adaptable speaker-dependent parameter, re-parametrisated by a function \( \xi : \mathbb{R} \to \mathbb{R}^{+} \), where \( s \) is the speaker. \( w_j^{s} \) is the \( j \)th column of the corresponding weight matrix \( W_l \in \mathbb{R}^{d_k \times d_h} \), \( b_j^{s} \) denotes the bias, \( \psi \) is the hidden unit activation function, and \( \circ \) denotes a Hadamard product.

In the original formulation of LHUC, for test-only adaptation, the speaker-dependent parameters \( \theta^{LHUC}_{LHUC} = \{ r_j^{s} \}_{j=1}^{L} \) and the speaker-independent parameters \( \theta_{SI} = \{ w_j^{s} \}_{j=1}^{L} \) were separately optimised. During training, the hidden units were estimated speaker-independently and were re-scaled by \( \xi(r^{s}) \), with the speaker dependent-amplitude parameters \( \theta^{LHUC}_{LHUC} \) estimated using adaptation data. In this work, we use speaker-specific information to learn hidden unit amplitudes during training.

This research was supported by EPSRC Programme Grant EP/I031022/1, Natural Speech Technology (NST). The NST research data collection may be accessed at http://datashare.is.ed.ac.uk/handle/10283/786.
example learning that a feature is harmful for some speakers but useful for others (look at Fig. 3 for an illustration). Likewise, similar properties (once learned) can be exploited during adaptation to unseen speakers resulting in better speaker-adapted models.

To perform SAT training with LHUC, we use the following objective:

$$L_{SA}(\theta_{SI}, \theta_{SD}) = -\sum_{t \in D} \log P(c_t|x_t^s; \theta_{SI}; \theta_{SD})$$

where \( s \) denotes the \( s \)th speaker, \( m_t \in \{0, s\} \) selects the SI or SD LHUC transforms from \( \theta_{SD} \in \{\theta_{LHUC}^0, \ldots, \theta_{LHUC}^L\} \) for each data-point separately (i.e. at the frame level, cf. Fig. 1) based on a Bernoulli distribution parametrised by \( \gamma \) hyper-parameter that determines the overall SI/SD data ratio, as follows:

$$k_t \sim \text{Bernoulli}(\gamma)$$

$$m_t = \begin{cases} s & \text{if } k_t = 0 \\ 0 & \text{if } k_t = 1 \end{cases}$$

3. EXPERIMENTAL SETUPS

We have evaluated SAT-LHUC using three different corpora: the TED talks corpus [29] following the IWSLT evaluation protocol (www.iwslt.org), the Switchboard corpus of conversational telephone speech [30] (ldc.upenn.edu) and the AMI meetings corpus [31][32] (corpus.amiproject.org). Unless explicitly stated otherwise, the models share a similar structure across the tasks – DNNs with 6 hidden layers (2,000 units in each) and a sigmoid non-linearity. The output logistic regression layer models the distribution of context-dependent clustered tied states [33]. The features are presented in 11 (±5) frame long context windows.

For TED we follow the recipe described in [34]. In this work however, compared to [34][16], our systems benefit from better language models developed for our IWSLT–2014 systems [15]: in particular, we rescroe using a 4-gram language model estimated from 751 million words. The baseline TED AMs are trained on unadapted PLP features with first and second temporal derivatives. We report the results on test2010 and test2013 sets. The latter is more challenging due to larger speaker variability as well as the need for automatic segmentation.

Fig. 1. Schematic of SAT-LHUC training.

In case of Switchboard (SWBD) we use the Kaldi GMM recipe [36][37], using Switchboard-1 Release 2 (LDC97S62). Our baseline unadapted acoustic models were trained on either MFCC or LDA/MLLT features. The results are reported on the full Hub5 00 set (LDC2002S09) to which we will refer as eval2000.

For AMI, we follow the Kaldi GMM recipe described in [38], which is using the so called AMI Full-ASR split on train, dev and eval sets. On this corpus we also train a separate set of models using mel-filter-bank (FBANK) features for which FMLLR transforms cannot be easily obtained, and as such, LHUC makes an interesting adaptation alternative.

The SAT related statistics for each of the above corpora are given in Table 1. Note, in this work we adapt to the headset or the side of a conversation, rather than the actual speaker: hence the number of clusters (or estimated transforms) during training can differ from the number of speakers.

Fig. 2. Example illustration on how LHUC performs adaptation (best view in color). A “bump” model with two hidden units can approximate “bump” functions (top). To learn function \( f_2 \) given training data \( f_1 \) (middle), we splice two “bump” functions together (4 hidden units, one input/output) to learn an approximation of function \( f_2 \). Let us assume that we want to adapt to \( f_2 \) using LHUC scalers. We plot the model optimised to \( f_1 \) and adapted to \( f_2 \) by adjusting only LHUC parameters (bottom).

Fig. 3. Example illustration showing how SAT-LHUC can improve a learned representation. Assume we want to approximate both \( f_1 \) and \( f_2 \) with the similarly constrained (4 hidden units) model from Fig. 2. Again, it is possible with two sets of SAT-LHUC parameters for \( f_1 \) and \( f_2 \).

In Table 1, we list the number of speakers, the number of training and test clusters, the number of training and test speakers, and the number of training and test sentences.
In the second experiment we investigated whether the inferior adaptation results for $\gamma < 0.3$ were caused by differences in learned representations or by lower quality adaptation targets. We used the adaptation targets of the ‘Baseline SI’ model (28.9% WER) and adapted SAT-LHUC models trained with $\gamma \in \{0.05, 0.1, 0.3, 0.5, 0.7, 0.1\}$ on 30-hour-TED. The results (Fig. 4(a)) indicate that the reason for lower adaptation accuracies (compared to $\gamma = 0.5$ system) was mostly due to less accurate adaptation targets. Adapting the $\gamma = 0.3$ model with the ‘Baseline SI’ targets reduces the WERs of $\gamma = 0.3$ system to 22.6% (from 23.2%) – 2.5% absolute lower when compared the baseline SI LHUC system (25.1%) (both systems used the same adaptation targets) and 0.1% absolute lower than the best $\gamma = 0.5$ system. This further strengthens our claim that the SAT-LHUC models indeed learn a better and more tunable speaker-dependent representation, but its use is somehow limited by a necessary trade-off of managing a good SI first-pass representation.

Finally, we investigated whether the inferior adaptation results for $\gamma < 0.3$ were caused by differences in learned representations or by lower quality adaptation targets. We used the adaptation targets of the ‘Baseline SI’ model (28.9% WER) and adapted SAT-LHUC models trained with $\gamma \in \{0.05, 0.1, 0.3, 0.5, 0.7, 0.1\}$ on 30-hour-TED. The results (Fig. 4(a)) indicate that the reason for lower adaptation accuracies (compared to $\gamma = 0.5$ system) was mostly due to less accurate adaptation targets. Adapting the $\gamma = 0.3$ model with the ‘Baseline SI’ targets reduces the WERs of $\gamma = 0.3$ system to 22.6% (from 23.2%) – 2.5% absolute lower when compared the baseline SI LHUC system (25.1%) (both systems used the same adaptation targets) and 0.1% absolute lower than the best $\gamma = 0.5$ system. This further strengthens our claim that the SAT-LHUC models indeed learn a better and more tunable speaker-dependent representation, but its use is somehow limited by a necessary trade-off of managing a good SI first-pass representation.
### Table 3. WER(%) and relative WER change (WERR)(%) on Switchboard Hub00. Feature-transform (FT) denotes fMLLR transforms.

<table>
<thead>
<tr>
<th>System</th>
<th>Training</th>
<th>Decoding</th>
<th>Features</th>
<th>Hub5'00</th>
<th>SWB</th>
<th>CHE</th>
<th>TOTAL</th>
<th>WERR (%)</th>
<th>Baseline Sys. ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline speaker-independent models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>SI</td>
<td>SI</td>
<td>MFCC</td>
<td>15.8</td>
<td>28.4</td>
<td>22.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>SI</td>
<td>SI</td>
<td>LDA/MLLT</td>
<td>15.2</td>
<td>28.2</td>
<td>21.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline speaker-adapted systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>SI</td>
<td>LHUC</td>
<td>MFCC</td>
<td>15.4</td>
<td>27.0</td>
<td>21.2</td>
<td>-4.5</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>SI</td>
<td>LHUC</td>
<td>LDA/MLLT</td>
<td>14.7</td>
<td>26.6</td>
<td>20.7</td>
<td>-4.6</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>SAT-F</td>
<td>FT</td>
<td>LDA/MLLT</td>
<td>14.2</td>
<td>26.2</td>
<td>20.2</td>
<td>-7.0</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>SAT-F</td>
<td>FT+LHUC</td>
<td>LDA/MLLT</td>
<td>14.2</td>
<td>25.6</td>
<td>19.9</td>
<td>-1.5</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>SAT Trained</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>SAT-LHUC</td>
<td>LHUC</td>
<td>MFCC</td>
<td>14.8</td>
<td>26.5</td>
<td>20.7</td>
<td>-6.3 / -2.4</td>
<td>A / C</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>SAT-LHUC</td>
<td>LHUC</td>
<td>LDA/MLLT</td>
<td>14.6</td>
<td>25.9</td>
<td>20.3</td>
<td>-6.5 / -1.9</td>
<td>B / D</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>SAT-F-T</td>
<td>FT+LHUC</td>
<td>LDA/MLLT</td>
<td>14.1</td>
<td>25.6</td>
<td>19.9</td>
<td>-0.0</td>
<td>F</td>
<td></td>
</tr>
</tbody>
</table>

The adaptation results of the SAT-LHUC model are given in Table 3 in row H (20.3%) where we almost match the SAT fMLLR baseline (20.2). We also observe that LHUC performs relatively better under more mismatched conditions – here Callhome (CHE) subset of eval2000 – similar to what was found on TED. Note, we train two sets of models, one on MFCC features to stay compatible with test-only adaptation techniques reported in [7] as well as linear discriminant analysis (LDA) features based on which Kaldi SWBD recipe [16] estimates fMLLR transforms - which form our baseline for the SAT training.

### Table 4. WER(%) on AMI.

<table>
<thead>
<tr>
<th>System</th>
<th>Training</th>
<th>Decoding</th>
<th>Features</th>
<th>dev</th>
<th>eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline speaker-independent systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SI</td>
<td>SI</td>
<td>FBANK</td>
<td></td>
<td>26.5</td>
<td>29.1</td>
</tr>
<tr>
<td>SAT-LHUC</td>
<td>SI</td>
<td>FBANK</td>
<td></td>
<td>26.3</td>
<td>28.9</td>
</tr>
<tr>
<td>Speaker-adapted systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SI</td>
<td>LHUC</td>
<td>FBANK</td>
<td></td>
<td>25.6</td>
<td>27.1</td>
</tr>
<tr>
<td>SAT-LHUC</td>
<td>LHUC</td>
<td>FBANK</td>
<td></td>
<td>24.9</td>
<td>26.1</td>
</tr>
<tr>
<td>SI</td>
<td>FT</td>
<td>FMLLR</td>
<td></td>
<td>26.2</td>
<td>27.3</td>
</tr>
<tr>
<td>SAT-F-T</td>
<td>FT+LHUC</td>
<td>FMLLR</td>
<td></td>
<td>25.6</td>
<td>26.2</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

We have proposed SAT-LHUC, an effective speaker adaptive training extension to the LHUC adaptation technique. SAT-LHUC does not require any auxiliary models or additional SAT training stages on top of the SI model to be effective, though it can be easily combined with other adaptation methods to bring further gains. The standalone variant is probably the simplest SAT approach proposed to date. This work is further extended in [42], in the future we plan to evaluate whether the proposed form of SAT remains effective with other types of non-linearities (as is the case for LHUC adaptation [16]), and an extension to sequence discriminative training [43].
6. REFERENCES


[26] C Wu and M Gales, “Multi-basis adaptive neural network for rapid adaptation in speech recognition,” in Proc ICASSP. 2015, IEEE.


