Speaker Normalization through Constrained MLLR Based Transforms

D. Giuliani+, M. Gerosa++ and F. Brugnara+

(+) Centro per la Ricerca Scientifica e Tecnologica, 38050 Panté di Povo, Trento, Italy
(++) University of Trento, 38050 Panté di Povo, Trento, Italy

Abstract

In this paper, a novel speaker normalization method is presented and compared to a well known vocal tract length normalization method. With this method, acoustic observations of training and testing speakers are mapped into a normalized acoustic space through speaker-specific transformations with the aim of reducing inter-speaker acoustic variability. For each speaker, an affine transformation is estimated with the goal of reducing the mismatch between the acoustic data of the speaker and a set of target hidden Markov models. This transformation is estimated through constrained maximum likelihood linear regression and then applied to map the acoustic observations of the speaker into the normalized acoustic space.

Recognition experiments made use of two corpora, the first one consisting of adults’ speech, the second one consisting of children’s speech. Performing training and recognition with normalized data resulted in a consistent reduction of the word error rate with respect to the baseline systems trained on unnormalized data. In addition, the novel method always performed better than the reference vocal tract length normalization method.

1. Introduction

It is well known that inter-speaker acoustic variability is one of the major sources of error in automatic speech recognition. Typical examples of sources of acoustic variations among speakers are the anatomical characteristics (e.g., vocal tract length, dimension of mouth and nasal cavities) and the speaking habits (e.g., accent, dialect and speaking rate). A lot of research effort has been devoted to investigate effective techniques for adapting an existing recognition system to each new speaker by exploiting a small amount of speech data collected from the speaker herself, with the aim of achieving recognition performance comparable to those of a speaker dependent recognition system [1, 2, 3].

In this work speaker adaptive acoustic modeling [4] is investigated by using two speaker normalization methods and their combination: the first one is a method for vocal tract length normalization (VTLN) [5, 6], the second one is a novel speaker normalization method based on constrained maximum likelihood linear regression (MLLR) [2, 3]. With these methods a suitable transformation has to be determined and applied for each speaker. For the VTLN method, the only parameter of such a transformation is the frequency scaling factor, to be applied for piece-wise linear frequency warping. In case of the constrained MLLR based speaker normalization (CMLSN) method, the parameters are the matrix and offset vector of the affine transformation to be applied for acoustic observation mapping. In both methods transformation parameters are estimated with the aim of increasing the matching score of the speaker’s data with a set of target hidden Markov models (HMMs). Estimation procedures require word transcriptions of speaker’s utterances.

Recognition experiments made use of two corpora, the Wall Street Journal (WSJ) corpus consisting of adults’ speech and the ChildIt corpus consisting of Italian children’s speech. Children exhibit significantly more intra- and inter-speaker acoustic variability with respect to adults [7], therefore speaker normalization methods are potentially suitable for automatic recognition of children’s speech [7]. Results reported in this paper show that the proposed speaker normalization method leads to a consistent reduction of the word error rate with respect to baseline systems for adults and children. Furthermore, it outperforms the VTLN method adopted in this work as reference for speaker normalization.

2. Speaker Normalization

Speaker normalization attempts to reduce inter-speaker acoustic variability induced by the different characteristics of the speakers. A typical speaker normalization scheme is one in which the speech data of each training and testing speaker are mapped into a normalized acoustic space through a speaker-specific transformation [8].

Training continuous density HMMs on normalized speech data should result in an improved modeling of phonetically relevant acoustic variations with output distributions showing reduced variances. In addition, performing recognition on normalized data should result in recognition performance less sensitive to the speaker’s characteristics.

In the following Sections the VTLN method adopted in this work is first introduced and then the novel speaker normalization method is described.

2.1. VTLN

VTLN aims to reduce inter-speaker acoustic variability due to vocal tract length variation among speakers [5]. Different VTLN methods have been proposed for warping the frequency axis of the speech spectrum to compensate for the vocal tract length of each training and testing speaker [5, 6]. With these methods, performing training and recognition in a “normalized” acoustic space allowed to improve performance of speaker-independent systems based on HMMs.

The training and recognition procedures adopted for implementing VTLN in this work follow closely those proposed in [6].
During training, for each speaker an optimal frequency scaling factor is estimated and applied to her/his data. A grid search over a possible set of frequency scaling factors is carried out to select the scaling factor which maximizes the likelihood of the speaker’s data with respect to a set of speaker independent triphone HMMs having a single Gaussian density per state and trained on unnormalized data. Word transcriptions of speaker’s utterances are required for this frequency scaling factor selection. A set of triphone HMMs are then trained on normalized data.

During recognition, word transcriptions are provided by a preliminary decoding step of unnormalized test utterances with the baseline system exploiting acoustic models trained on unnormalized data. For each test speaker, HMMs trained on normalized data are first exploited for maximum likelihood selection of the frequency scaling factor, then they are used for decoding with speech data normalized by applying the speaker-specific frequency scaling factor.

In this work, the warping process is implemented by changing the spacing and the width of the filters in the mel filter-bank while maintaining the speech spectrum unchanged [5]. In particular, a piece-wise linear warping function of the frequency axis of the filters is assumed to cope with the problem of accommodating filters near the upper band-edge [5].

## 2.2. Constrained MLLR based Speaker Normalization

Instead of warping the power spectrum during signal analysis, speaker normalization can be performed transforming the acoustic observation vectors by means of an affine transformation.

In [3] it is shown that mean \( \mu \) and variance \( \Sigma \) of a Gaussian density \( \mathcal{N}(x; \mu, \Sigma) \) associated with an HMM state can be adapted by means of an affine transformation, estimated in the MLLR framework, in the following way:

\[
\hat{\mu} = \tilde{A}\mu + \tilde{b}, \quad \hat{\Sigma} = \tilde{A}\Sigma\tilde{A}^\top
\]

where \( \tilde{A} \) and \( \tilde{b} \) represent the matrix and the offset vector of the so called constrained model-space transformation [3]. The term constrained denotes that the same matrix is applied to transform mean and variance. When a single transformation is used for adapting all the Gaussian densities in the recognition system, constrained MLLR adaptation can be implemented by transforming acoustic observations, thanks to the following identity:

\[
\mathcal{N}(Ax + b; \mu, \Sigma) = |A|^{-1} \mathcal{N}(x; A^{-1}(\mu - b), A^{-1}\Sigma A^{-\top}).
\]

In this work, we are interested in the feature-space transformation, to be applied to acoustic observations, represented by \( A \) and \( b \) which are related to \( \tilde{A} \) and \( \tilde{b} \) by: \( \tilde{A} = A^{-1}, \tilde{b} = -A^{-1}b. \)

For each speaker, normalization is performed with respect to a set of target HMMs trained on unnormalized data. Estimation of \( A, b \) is carried out following the Expectation-Maximization (EM) approach which requires to maximize the following auxiliary function in order to increase the likelihood of the transformed data:

\[
Q = \epsilon - \frac{1}{2} \sum_{\omega \in O} \sum_{g=1}^G \sum_{t=1}^{T(\omega)} \gamma_g(t) \left(- \log(|A|^2) + (A\omega(t) + b - \mu_g)^\top \Sigma_g^{-1} (A\omega(t) + b - \mu_g)\right)
\]

where \( O \) denotes the set of available utterances, \( g \) denotes a Gaussian density of the target HMMs and \( t \) denotes a time frame. \( \epsilon \) is a constant that does not depend on \( A \) and \( b \), and \( \gamma_g(t) \) represents the conditional probability of Gaussian density \( g \) at time \( t \) given the observation sequence and the current set of model parameters. A re-estimation procedure for \( A \) and \( b \) is proposed in [3] under the assumption of diagonal covariance matrices.

In this work, normalization is carried out with respect to a set of triphone HMMs having a single Gaussian density, with diagonal covariance matrix, per state and trained on unnormalized data. These models are used as target models during both training and recognition.

Concerning the structure of the target models, we argue that using many Gaussian densities per HMM state would make the constrained MLLR transformation less effective for our purposes, as inter-speaker variability would be already compensated by model parameters. This was confirmed by preliminary experiments using target models with 2 and 4 Gaussians per state, and has been also remarked upon in [6], in the context of maximum likelihood selection of the frequency scaling factor for VTLN. In addition, a full transformation matrix is adopted for mapping all the components (i.e. static and dynamic features) of observation vectors.

The training and decoding procedures resemble those for VTLN defined in [6]. The procedure for training speaker independent HMMs in a normalized feature space is the following:

1. Generate and train a set of tied-state triphone HMMs \( \lambda \) with a single Gaussian density per state by exploiting unnormalized training data
2. For each training speaker \( r = 1, ..., R \):
   - Estimate transformation parameters \( A_r \) and \( b_r \) in order to maximize the auxiliary function of Eq. (1) given the model set \( \lambda \) and the set of speaker’s utterances \( O_r \) with the corresponding set of word transcriptions \( W_r \)
   - Transform and store acoustic observations
     \[
     \tilde{O}_r = \{[\tilde{a}(t) = A_r \omega(t) + b_r]_{t=1}^{T(\omega)} | \omega \in O_r \}
     \]
3. Generate and train a new set of HMMs \( \tilde{\lambda} \) exploiting normalized training data.

The decoding procedure for a given test speaker \( s \) is the following:

1. Generate word hypotheses \( \tilde{W}_s \) for the speaker’s utterances \( O_s \) by decoding with baseline models \( \lambda \) trained on unnormalized data
2. Estimate transformation parameters \( A_s \) and \( b_s \) in order to maximize the auxiliary function of Eq. (1) given the model set \( \lambda \) and the set of speaker’s utterances \( O_s \) with the corresponding set of word transcriptions \( W_s \)
3. Transform speaker’s utterances
   \[
   \tilde{O}_s = \{[\tilde{a}(t) = A_s \omega(t) + b_s]_{t=1}^{T(\omega)} | \omega \in O_s \}
   \]
4. Decode normalized utterances \( \tilde{O}_s \) using the \( \tilde{\lambda} \) model set trained on normalized data.

With respect to VTLN, a possible advantage in transforming acoustic observations with an affine transformation estimated in this way is that no assumption is done about the nature of sources of speaker individualities. For example spectral differences between speakers could be due to linguistic differences (e.g. accent or dialect) in addition to vocal tract variations. Moreover, this approach is also suitable for normalizing
acoustic data with respect to other sources of acoustic variability like microphones, transmission channels and acquisition environments. A noticeable case in which the method is naturally applied is that of combined speaker and channel mismatch. On the other hand, a possible limitation is that nonlinear spectral variation can be hardly compensated for by a affine transformation.

The proposed method uses the same kind of transformation adopted in [3] for implementing constrained MLLR based speaker adaptive training (SAT). However, the two methods differ in the model set used to perform the transformation estimation. While in [3] estimation of speaker-specific transformations is carried out with respect to the complex model set to be used in recognition, here transformation estimation is carried out with respect to a set of target models, having a single Gaussian density per state and trained on unnormalized data, with the aim of forcing a normalization on the data. This target model set is fixed, so that the speaker-specific transformations are computed and applied only once before starting the training procedure. In addition, in contrast with SAT where usually baseline SI HMMs are used as seed models, here a new set of models is generated from scratch (including creation of a phonetic decision tree for HMM states tying) and trained exploiting the normalized data. During the decoding stage transformation estimation is still carried out with respect to the target models set before decoding with HMMs trained with normalized data. We point out that HMM adaptation can be still performed after data normalization.

3. Recognition Experiments

Two speech corpora were employed for experiments, the WSJ0+1 corpus and the Italian corpus Childl consisting of read children’s speech.

For recognition experiments we used the ITC-irst HMM package employing state-tied, cross-word triphone HMMs. Output distributions associated with HMM states were modeled with mixture densities having up to 8 Gaussian components with diagonal covariance matrices.

Each speech frame was parameterized into 12 mel frequency cepstral coefficients (MFCC) and log-energy. These coefficients plus their first and second order time derivatives were combined to form 39-dimensional observation vectors. Cepstral mean subtraction was performed on an utterance by utterance basis while log-energy was normalized with respect to the maximum value in the utterance.

3.1. Experiments with the WSJ Corpus

System training was performed with WSJ0+1 SI-284 training set. 57 hours of speech data were used after shortening long initial and final segments of silence.

Acoustic models were gender independent cross word triphone HMMs. A phonetic decision tree was used for tying HMM states. The resulting number of tied-states for the baseline models, trained on unnormalized data, was 8874. A similar number of tied-states was obtained when training on normalized data.

Three sets of evaluation data were used. The ARPA 1993 H1 evaluation consists of 213 sentences from 10 speakers. The ARPA 1993 H2 evaluation set consists of 215 sentences from 10 different speakers. Recordings in the H1 and H2 evaluation sets were performed with the same microphone, a head-worn close talk microphone, used for the SI-284 training set. In order to test performance under unmatched acoustic conditions, the ARPA 1993 S5 evaluation set was employed. S5 consists of recordings from the secondary microphone used when performing recordings in H2. For each speaker, one of 10 alternative microphones was used. H2 and S5 therefore contain the same utterances recorded with different microphones.

The ARPA November 1993 20k word trigram language model was used in all the recognition experiments carried out on H1, H2 and S5 evaluation sets.

Table 1 reports recognition results, in terms of word error rate (WER), obtained on the WSJ 1993 H1, H2 and S5 evaluation sets by using the baseline system, applying vocal tract length normalization only during recognition, and performing unsupervised static speaker adaptation. After some preliminary experiments we chose to perform unsupervised static speaker adaptation by adapting parameters of Gaussian densities through MLLR transformations of the means. Two regression classes were defined and the associated full transformation matrices were estimated through two MLLR iterations. Both for MLLR and VTLN, word transcriptions for test utterances of each speaker were provided by a preliminary decoding step carried out with the baseline system.

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Baseline</th>
<th>VTLN</th>
<th>MLLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>12.9</td>
<td>13.1</td>
<td>11.6</td>
</tr>
<tr>
<td>H2</td>
<td>6.8</td>
<td>6.9</td>
<td>5.8</td>
</tr>
<tr>
<td>S5</td>
<td>18.2</td>
<td>16.8</td>
<td>8.4</td>
</tr>
</tbody>
</table>

Table 1: Recognition results (WER) on the WSJ 1993 H1, H2 and S5 evaluation sets for the baseline system (Baseline), by performing vocal tract length normalization only during the recognition stage (VTLN), and by performing MLLR speaker adaptation (MLLR). In round brackets the WER relative reduction (%) achieved with respect to the baseline system is reported.

On the H1 evaluation set the baseline system provides a 12.9% WER, which is comparable with results reported in the literature for this task. By considering all three evaluation sets, we can conclude that applying VTLN only during the recognition stage yields no or marginal performance improvement, while significant performance improvements are obtained by unsupervised static speaker adaptation based on MLLR.

<table>
<thead>
<tr>
<th>Test Set</th>
<th>CMLSN</th>
<th>VTLN</th>
<th>VTLN+CMLSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>11.0</td>
<td>12.0</td>
<td>10.5</td>
</tr>
<tr>
<td>H2</td>
<td>5.6</td>
<td>6.4</td>
<td>5.5</td>
</tr>
<tr>
<td>S5</td>
<td>11.2</td>
<td>15.5</td>
<td>10.2</td>
</tr>
</tbody>
</table>

Table 2: Recognition results (WER) on the WSJ 1993 H1, H2 and S5 evaluation sets by performing speaker normalization with different methods.

Recognition results obtained by using different speaker normalization methods are shown in Table 2. In addition to results for CMLSN and VTLN methods, Table 2 reports results obtained by combining VTLN with CMLSN. By considering the H1 evaluation set, CMLSN allows a 14.7% WER relative reduction, better than results achieved with both the VTLN method, 7.0% WER relative reduction, and MLLR adaptation of the means of baseline models, 10% WER relative reduction (see Table 1).

Results obtained on the H2 and S5 evaluation sets allow to verify the effectiveness of CMLSN under unmatched train-
ing and testing acoustic conditions. In case of matched conditions a WER relative reduction of 17.6% is achieved while under unmatched conditions a WER relative reduction of 38.4% is achieved. By considering the three evaluation sets, CMLSN always outperforms VTLN. However, when CMLSN is compared with MLLR speaker adaptation of baseline models (see Table 1), we note that in the case of matched training and testing acoustic conditions (H1 and H2) CMLSN outperforms MLLR speaker adaptation, while in the case of unmatched training and testing acoustic conditions (S5) MLLR speaker adaptation performs tangibly better. This can be explained by the fact that CMLSN performs speech data normalization toward a set of target HMMs trained on clean speech and having a single Gaussian per state, while MLLR speaker adaptation adapts the acoustic models to be used for the final decoding step.

Experiments carried out with the purpose of exploring the potential of combining VTLN with the CMLSN method, VTLN+CMLSN column in Table 2, always result in a performance improvement with respect to applying one of the two methods individually. However, on the evaluation set S5 performance is still lower than that achieved with MLLR speaker adaptation of the baseline models.

### 3.2. Experiments with the ChildIt Corpus

The ChildIt corpus consists of Italian read sentences collected from 171 children (86 male and 85 female) aged between 7 and 13. This corpus was developed at ITC-irst. Recordings took place at school, usually in the computer room or in the library. Each child was asked to read a set of sentences prepared according to her/his grade.

System training was performed by using 8 hours of speech collected from 129 speakers. Acoustic models were gender-independent cross word triphones. A phonetic decision tree was used for clustering HMM states. The number of tied-states for the baseline models, trained on unnormalized data, was 1400. A similar number of tied-states was obtained when training on normalized data.

The evaluation set consists of 2.5 hours of speech collected from 42 children, evenly distributed per age and gender, and a 11k word trigram language model was used for recognition experiments. The dictionary was composed of words occurring in both training and test sets. The language model was estimated on a text corpus consisting of articles taken from Italian newspapers.

In Table 3 recognition results obtained with the baseline system, by unsupervised static speaker adaptation of baseline models and by speaker normalization methods are reported. The baseline system provides a 15.1% WER, while MLLR speaker adaptation yields a 15.2% WER relative reduction. We point out that the same performance improvement is obtained by transforming acoustic observations by means of an affine transformation estimated through constrained MLLR. WER relative reductions of 24.5% and 11.2% are achieved with CMLSN and VTLN normalization methods, respectively. Combining VTLN and CMLSN doesn’t lead to a performance improvement with respect to the CMLSN alone. Comparing results achieved for adults and children, speaker normalization methods seem more effective for children. However, we must point out that the baseline system for adults is trained by using more data and a larger number of training speakers with respect to the baseline system for children. Therefore, the baseline system for adults is likely to be more robust with respect to inter-speaker acoustic variations. This is supported by the fact that applying VTLN only during recognition for adults doesn’t result in any performance improvements (see Table 1) while for children it results in a 9.2% WER relative reduction.

<table>
<thead>
<tr>
<th>Baseline</th>
<th>MLLR</th>
<th>CMLSN</th>
<th>VTLN</th>
<th>VTLN+CMLSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.1</td>
<td>12.8 (15.2)</td>
<td>11.4 (24.5)</td>
<td>13.4 (11.2)</td>
<td>11.5 (23.8)</td>
</tr>
</tbody>
</table>

Table 3: Recognition results (WER) on the ChildIt evaluation set for the baseline system, by performing MLLR speaker adaptation of baseline models and by speaker normalization methods.

### 4. Conclusions

In this paper we have proposed a speaker normalization method to deal with inter-speaker acoustic variability and we have compared it to a well known vocal tract length normalization method. Recognition results have shown that the proposed method systematically outperforms the reference normalization method based on VTLN.

Furthermore, under matched training and testing acoustic conditions, the proposed method performs better than MLLR adaptation of baseline acoustic models trained on unnormalized data. On the other hand, under unmatched training and testing acoustic conditions, MLLR adaptation of baseline models ensures better performance.

Future work will be devoted to investigate the potential of combining the proposed normalization method with MLLR acoustic model adaptation.

### 5. References


