Invariant Integration Features Combined with Speaker-Adaptation Methods

Florian Müller and Alfred Mertins

Institute for Signal Processing
University of Lübeck, Germany
{mueller,mertins}@isip.uni-luebeck.de

Abstract

Speaker-normalization and -adaptation methods are essential components of state-of-the-art speech recognition systems nowadays. Recently, so-called *invariant integration features* were presented which are motivated by the theory of invariants. While it was shown that the integration features outperform MFCCs when used with a basic monophone recognition system, it was left open, if their benefits still can be observed when a more sophisticated recognition system with speaker-normalization and/or speaker-adaptation components is used. This work investigates the combination of the integration features with standard speaker-normalization and -adaptation methods. We show that the integration features benefit from adaptation methods and significantly outperform MFCCs in matching, as well as in mismatching training-test conditions.
Index Terms: Speaker independency, invariant integration, speaker normalization, speaker adaptation

1. Introduction

An essential component of state-of-the-art *automatic speech recognition* (ASR) systems is the adaptation of the system to the characteristics of the speech signals of each speaker. The adaptation methods work at different stages of an ASR system. With respect to the feature extraction, a common approach is the normalization of the mean and variance of the speech signal's parametric representation [1]. The *vocal-tract length* (VTL) as an inter-speaker variability [2] causes the formant frequencies approximately to be linearly scaled. Either scaling the filter-bank frequency centers, or scaling the time-frequency (TF) representation along the frequency axis, respectively, can be done in a *maximum-likelihood* (ML) fashion [3, 4]. This widely used procedure is known as *vocal-tract length normalization* (VTLN). With respect to the acoustic models, different adaptation methods have become standard components in commercial ASR systems. The general idea is to adapt the model set such that it matches more closely the features of the current speaker. Constrained and unconstrained *maximum-likelihood linear regression* ((C)MLLR) [5, 6] are most-widely used approaches. Though it was shown in [7] that VTLN can be seen as a special case of CMLLR, in practice the combination of both methods within one system often increases its accuracy further.

Besides the *mean and variance normalization* (MVN) as mentioned above, a third group of methods tries to extract features that are independent of the speaker's variabilities while keeping the important linguistic information. The different approaches usually rely on an invariance property of a certain type of transformation. Because the spectral effects of different VTLs can approximately be described by a scaling along a linear frequency axis, the scale transformation [8] was investigated for its applicability in speech recognition [9, 10]. Refinements towards real-time applicability of this type of transformation have been presented in [11].

The application of auditory scales like the mel [12] or the ERB [13] scale approximately maps the scaling to translation along the frequency axis of TF representations. This effect was used for translation-based VTLN [14], as well as for *Gaussian-mixture-model* (GMM) based features [15]. Recently, different types of translation-invariant transformations were investigated for their applicability in the field of speech recognition. Correlation-based features were proposed in [16]. Methods based on invariant cyclic transformations were considered in [17, 18]. So-called *invariant integration features* (IIFs) were introduced in [19]. The transformations that are used within these feature extraction methods are mathematically well founded and were successfully used in the field of image analysis.

Using a basic monophone recognition system without adaptation, the work in [19] primarily introduced the IIFs and showed that the IIFs can outperform the *mel frequency cepstral coefficients* (MFCCs). The present paper uses a contextually enhanced definition of the IIFs and examines their performance in a more sophisticated ASR system that makes use of speaker-normalization and -adaptation methods. The number of subbands of the filterbanks that are used for the computation of features like the MFCCs is usually between 24 and 30 bands. In contrast, the experiments in [19] used a filterbank with 90 subbands. This was done to allow for a sufficiently high spectral resolution. However, the computational complexity also increases with increasing number of filters. To reduce costs, part of this work investigates IIFs based on a TF representation with 24 subbands.

The next section briefly describes the IIFs and their contextual expansion. Since IIFs were introduced to provide a technique for speaker-independent feature extraction the question may arise, why speaker-normalization techniques should improve the performance of the IIFs. This is also discussed in Section 2. The experimental setup and the results are described in Section 3. The paper is concluded in Section 4.

2. Invariant Integration Features

The basic motivation for the IIFs is the observation that spectral effects due to different VTLs are mapped to translations along the subbands-index space within the TF representation when an auditory scale is used. The IIFs were presented in [19]. A more detailed study was done in [20], in which the definition of the IIFs was contextually enhanced. In the following, a brief description of the IIFs is given.

Copyright © 2010 ISCA

2.1. Definition and Feature Selection

The first step in the feature-generation process is a filterbank analysis of the speech signal, using, for example, a gammatone filterbank. Let $y_k(n)$ denote the magnitudes of the obtained subband coefficients at the final frame rate, where n is the time and k the subband index, with $1 \leq n \leq N$ and $1 \leq k \leq K$. Given the indices vector $\boldsymbol{k} = (k_1, k_2, \ldots, k_M)$, $\boldsymbol{k} \in \mathbb{N}^M$, an integer exponent vector $\boldsymbol{l} = (l_1, l_2, \ldots, l_M)$, $l \in \mathbb{N}_0^M$, and a temporal offset vector $\boldsymbol{m} \in \mathbb{N}^M$, a (contextual) monomial $\hat{m}$ with M components is defined as

$$\hat{m}(n; w, \boldsymbol{k}, \boldsymbol{l}, \boldsymbol{m}) := \left[\prod_{i=1}^{M} y_{k_i+w}^{l_i}(n + m_i) \right]^{1/\gamma(\hat{m})}, \quad (1)$$

where $w \in \mathbb{N}_0$ is a spectral offset parameter that is used for the ease of notation in the following, and $\gamma(\hat{m})$ is the *order of a monomial* $\hat{m}$ defined as

$$\gamma(\hat{m}) := \sum_{i=1}^{M} l_i \,. \quad (2)$$

The term (2), which occurs in the exponent in (1), operates as a normalizing term with respect to the order of the monomials. Now, an IIF $A_{\hat{m}}(n)$ is defined as

$$A_{\hat{m}} := \frac{1}{2W+1} \sum_{w=-W}^{W} \hat{m}(n; w, \boldsymbol{k}, \boldsymbol{l}, \boldsymbol{m}). \quad (3)$$

In case of order one, the computation of an IIF is equivalent to the computation of an average of spectral values within a certain frequency range.

Apparently, the parameter space of the IIFs is quite large and choosing an appropriate set of features is non-trivial. In this work, the same feature-selection approach as in [19] is used. It is an iterative filter method that is based on a linear classifier [21].

2.2. Decreasing the Number of Filters

Apart from context the definition of an IIF as given in Equation (3) is equal to the one given in [19]. The motivation for using integer window sizes W originates from the theory of invariants for finite groups [22] and computational efficiency. In the context of speaker-independent speech recognition, a finite set of translations along the frequency axis is assumed when using an appropriate filterbank. To have a sufficiently high spectral resolution, an appropriate number of filters has to be used. Previous works used 60 to 90 filters in this context [14, 16, 19]. However, to keep the computational complexity low, a small number of filters for the spectral analysis is desirable. By interpolating the spectral values of a TF representation with a small number of subbands, the same IIF definition as given in Equation (3) can be used. The experimental part in Section 3 investigates the impact of different numbers of subbands in the TF representations and the use of interpolation in more detail.

2.3. Combining IIFs with Speaker-Adaptation Methods

The windows sizes W, i.e., the spectral integration ranges of the used IIFs, are in general chosen in such a way, that the integration does not take place over the full bandwidth of the TF representation. Thus, the IIFs are only invariant within a certain subband range, and applying a VTLN before computing IIFs could also improve the accuracy of the IIFs. In this context, we propose to apply a VTLN that shifts the spectral values (similar to [14]). Due to the local invariance property, however, the increase in accuracy is expected to be lower compared to MFCCs.

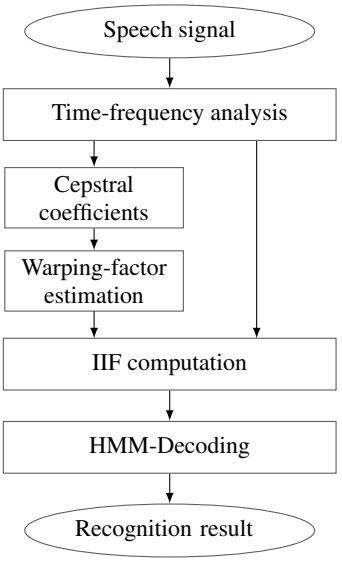

Figure 1: Combining VTLN with IIFs. Before computing the IIFs, a translational VTLN is applied to ML-estimate a translation factor for each speaker.

Figure 1 summarizes how VTLN is combined with IIFs in this work. At first, a TF analysis of the speech signal is computed. Here, an ERB scale is used in order to map the spectral scaling to translations along the subband index space. Following the general approach of VTLN as described in [4], a warping factor is estimated in a ML fashion with a grid search. During the VTLN, warped versions of the same TF representation are computed by shifting the spectral values along the frequency axis, and cepstral coefficients are computed subsequently. Finally, the IIFs are computed on basis of the ML-warped TF representations. When combined with VTLN, the feature selection for IIFs may also be conducted on the normalized TF representations.

In contrast to IIFs and VTLN, which are generally working on the feature-extraction stage, MLLR seeks to adapt the acoustic models in a ML fashion to the characteristics of each speaker. The combination of VTLN and MLLR has proven to be beneficial, e.g., [23], and both techniques are usually combined within state-of-the-art ASR systems. Because one may suspect that the use of normalization and adaptation methods may lead to a disappearing of the observed performance gains of the IIFs, the experimental part of this work investigates the effects of adaptation on the IIF performance.

3. Experiments

3.1. Data and Setup

Phone recognition experiments have been conducted on the TIMIT corpus [24] with a sampling rate of 16 kHz. It contains 6300 utterances read by 630 female and male adult speakers. By omitting the SA-sentences the training set consists of 3696 sentences and the test set consists of 1344 utterances. Following the standard approach for this corpus [25], the original 61 phonetic labels were collapsed to a set of 48 labels.

Two different training-testing scenarios were considered in the experiments; the matching scenario consisted of the standard training and test set. The mismatching scenario, however, used only the male utterances from the training set and only the female utterance from the test set. Thus, a mismatch between training and test conditions with respect to the mean VTL was simulated. Triphone context-dependent HMMs were trained as acoustic models. The number of *Gaussian-mixture model* (GMM) components of the output distributions was 16 for MFCC features. In contrast, it was found beneficial to use eight components in case of IIFs. All output distributions were modeled with diagonal covariance matrices. Decision-tree clustering was applied for tying triphone states. A bigram language model was used in all experiments.

For testing, the phonetic labels were further collapsed to 39 labels [25]. Standard MFCCs with 12 coefficients plus log-energy and first and second order derivatives were used for the acquisition of baseline accuracies. The filterbank used for the computation of the MFCCs had 24 filters. MVN was applied on all features.

For the TF representation for computing the IIFs, a gammatone filterbank based on an FFT-approach [26] was used. The filter's center frequencies were equally spaced along the ERB scale with a minimum frequency of 40 Hz and a maximum frequency of 8 kHz. A power-law compression with an exponent of 0.1 was applied on the spectral values of the TF representations. Experiments were conducted to compare the performance of IIFs for different numbers of subbands:

1. TF analysis with 24 filters,
2. TF analysis with 24 filters, afterwards interpolated to 110 spectral values, and
3. TF analysis with 110 filters.

Moreover, the performance of the considered feature types was investigated when combined with VTLN solely, with MLLR solely, and with both VTLN and MLLR. In case of MFCCs, scaling factors from 0.88 to 1.12 with a step size of 0.02 were used for the ML grid-search within the VTLN stage. In case of the IIFs, shifting factors from -1.5 to 1.5 with a step size of 0.25 were used for the 24-band TF analysis, and shifting factors from -8 to 8 with a step size of 1 were used for the 110-band TF analysis. When MLLR was applied, *speaker-adaptive training* (SAT) with CMLLR and speaker-adaptation with a combination of CMLLR and MLLR was used. A regression class tree with eight terminal nodes was employed.

3.2. Feature Selection

All experiments use 30 IIFs of order one. The choice of this order has been acquired in [20] and has proven to yield IIFs that outperform MFCCs in matching, as well as in mismatching scenarios. The contextual offsets in $\boldsymbol{m}$ are constrained to the interval $[-3, \ldots, 3]$, which corresponds to a maximum contextual interval of 80 ms. The maximum window size W is set to the number of subbands within the individual experiments. With these constraints, the feature selection as described in [19] was performed for 1500 iterations. The data used within the feature selection represented a matching scenario. Similar to the MFCCs, the log-energy together with first and second order derivatives were appended to the IIFs for the experiments. A *linear discriminant analysis* (LDA) followed by a *maximum-likelihood linear transform* (MLLT) [27] was applied to allow for diagonal covariance modeling of the IIFs.

Table 1: *Results of the experiments. MFCC- and IIF-based ASR systems with and without VTLN and/or MLLR are compared.*

#Subbands	Features	Adaptation	Accuracy [%] Scenario matching	mismatching
24	MFCC	-	72.16	54.42
		MLLR	75.18	66.98
		VTLN	73.27	69.75
		VTLN+MLLR	75.36	71.84
24	IIF	-	75.20	60.06
		MLLR	75.59	68.57
		VTLN	76.85	69.63
		VTLN+MLLR	77.16	**74.32**
24 → 110	IIF	-	75.24	61.51
		MLLR	75.97	69.38
		VTLN	77.12	70.79
		VTLN+MLLR	77.39	73.03
110	IIF	-	75.23	60.60
		MLLR	76.20	68.79
		VTLN	77.22	70.92
		VTLN+MLLR	**77.45**	72.71

3.3. Baseline Accuracies

As described above, baseline accuracies were computed with MFCCs combined with VTLN and/or MLLR. The results are shown in the upper part of Table 1. Within the table, the highest accuracies in each scenario are shown in bold.

As expected, it can be observed that the accuracy increases, when VTLN or MLLR is used, and the largest gain in accuracy is achieved when VTLN and MLLR are combined within the same system. In comparison, the benefits of normalization and/or adaptation are larger in the mismatching scenario than in the matching one.

3.4. Results for Invariant Features

The results for the IIFs are shown below the MFCC results in Table 1. The number of subbands used within the TF representation (c.f. Section 3.1) is indicated in the left column. Generally, it can be stated that the IIFs clearly benefit from additional speaker-normalization and -adaptation methods. While enhancing the MFCC-based ASR system with VTLN and MLLR increases the accuracy for the matching scenario by about three percentage points, the increase of accuracy of the IIF-based systems is about two percentage points for the matching scenario. For the mismatching scenario, the accuracy of the MFCC-based system is increased by about 17 percentage points. In contrast, the IIFs perform about twelve percentage points better when VTLN and MLLR is used. It can be observed that, in the matching scenario, all IIF accuracies are higher than the corresponding MFCC accuracies. In the mismatching scenario, the IIF combinations with VTLN and with VTLN plus MLLR yield in all cases higher accuracies than the corresponding results for the MFCCs. Remarkably, the IIFs without adaptation already perform as good as MFCCs together with VTLN+MLLR in the common matching scenario.

While using 110 subbands yields the highest accuracies within these experiments, the corresponding accuracies based on the TF representation with 24 subbands are only slightly lower. Comparing the results that are based on 24 subbands without interpolation with the results based on the interpolated TF representation, it can be observed that the latter leads to

higher accuracies in the matching case. However, the highest accuracy for the mismatching case was achieved with a 24-band TF representation without interpolation.

4. Discussion and Conclusion

In this work, we have shown that the recently presented *invariant integration features* (IIFs) can be combined with VTLN and/or MLLR without significantly increasing the complexity of the ASR system. Phone recognition experiments on TIMIT have shown that IIFs perform superior compared to MFCCs in both matching and mismatching scenarios. The superiority in the matching case is different to the observations in [19]. However, only IIFs of order one have been used in this work and the feature selection was conducted on data of a matching scenario. When using a small number of subbands for a TF representation (here 24 subbands) the accuracy of IIF-based systems decreases only slightly. We have shown that IIFs benefit from additional speaker-normalization and -adaptation methods like VTLN and/or MLLR. The benefits of the IIFs, that can be observed when used with a basic recognition system without any normalization/adaptation methods are still observable in combination with VTLN and/or MLLR.

Future work will be directed toward a more sophisticated feature selection algorithm and the use of physiologically motivated principles, which might lead to even more robust features. A software for computing the IIFs used in this work can be downloaded from `www.isip.uni-luebeck.de` under "Downloads".

5. Acknowledgements

This work has been supported by the German Research Foundation under Grant No. ME1170/2-1.

6. References

[1] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Handbook of Speech Processing*. Berlin, Germany: Springer, 2008.

[2] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: a review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, Oct.-Nov. 2007.

[3] L. Lee and R. C. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, Jan. 1998.

[4] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 6, pp. 415–426, Sept. 2002.

[5] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[6] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, Apr. 1998.

[7] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, Sept. 2005.

[8] L. Cohen, "The scale representation," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3275–3292, Dec. 1993.

[9] S. Umesh, L. Cohen, and D. Nelson, "Frequency-warping and speaker-normalization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'97)*, Apr. 1997, pp. 983–986.

[10] S. Umesh, L. Cohen, N. Marinovic, and D. J. Nelson, "Scale transform in speech analysis," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 1, pp. 40–45, Jan. 1999.

[11] A. D. Sena and D. Rocchesso, "A study on using the mellin transform for vowel recognition," in *Proc. Sound and Music Conf.*, Salerno, Italy, Nov. 2005.

[12] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[13] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception. Advanced Bioscience*, Y. Cazals, L. Demany, and K. Horner, Eds., vol. 83, Pergamon, Oxford, 1992, pp. 429–446.

[14] R. Sinha and S. Umesh, "Non-uniform scaling based speaker normalization," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'02)*, vol. 1, Orlando, USA, May 2002, pp. I–589 – I–592.

[15] J. J. Monaghan, C. Feldbauer, T. C. Walters, and R. D. Patterson, "Low-dimensional, auditory feature vectors that improve vocal-tract-length normalization in automatic speech recognition," *J. Acoustical Society of America*, vol. 123, no. 5, pp. 3066–3066, Jul. 2008.

[16] A. Mertins and J. Rademacher, "Frequency-warping invariant features for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. V, Toulouse, France, May 2006, pp. 1025–1028.

[17] F. Müller, E. Belilovsky, and A. Mertins, "Generalized cyclic transformations in speaker-independent speech recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Merano, Italy, Dec. 2009, pp. 211–215.

[18] F. Müller and A. Mertins, "Nonlinear translation-invariant transformations for speaker-independent speech recognition," in *Advances in Nonlinear Speech Processing*, ser. LNAI, J. Sole-Casals and V. Zaiats, Eds., vol. 5933. Heidelberg, Germany: Springer, Feb. 2010, pp. 111–119.

[19] ——, "Invariant-integration method for robust feature extraction in speaker-independent speech recognition," in *Proc. Int. Conf. Spoken Language Processing (Interspeech 2009-ICSLP)*, Brighton, UK, Sept. 2009, pp. 2975–2978.

[20] ——, "Contextual invariant-integration features for improved speaker-independent speech recognition," *Speech Communication*, submitted.

[21] T. Gramss, "Word recognition with the feature finding neural network (FFNN)," in *Proc. IEEE Workshop Neural Networks for Signal Processing*, Princeton, NJ, USA, Oct. 1991, pp. 289–298.

[22] E. Noether, "Der Endlichkeitssatz der Invarianten endlicher Gruppen," *Mathematische Annalen*, vol. 77, no. 1, pp. 89–92, Mar. 1915.

[23] J. Lööf, M. Bisani, C. Gollan, G. Heigold, B. Hoffmeister, C. Plahl, R. Schlüter, and H. Ney, "The 2006 RWTH parliamentary speeches transcription system," in *Proc. Interspeech 2006 - ICSLP*, Pittsburgh, USA, Sept. 2006, pp. 105–108.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *DARPA TIMIT acoustic phonetic speech corpus*, Linguistic Data Consortium, Philadelphia, 1993.

[25] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[26] D. P. W. Ellis, "Gammatone-like spectrograms," web resource: http://www.ee.columbia.edu/dpwe/resources/matlab/gammatonegram, Jun. 2009.

[27] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'98)*, Seattle, USA, May 1998, pp. 661–664.