Noise Robust Speaker-Independent Speech Recognition with Invariant-Integration Features Using Power-Bias Subtraction

Florian Müller and Alfred Mertins

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

Abstract

This paper presents new results about the robustness of invariant-integration features (IIF) in noisy conditions. Furthermore, it is shown that a feature-enhancement method known as "power-bias subtraction" for noisy conditions can be combined with the IIF approach to improve its performance in noisy environments while keeping the robustness of the IIFs to mismatching vocal-tract length training-testing conditions. Results of experiments with training on clean speech only as well as experiments with matched-condition training are presented.

Index Terms: speech recognition, speaker independency, noise robustness, invariant integration, power normalization

1. Introduction

Automatic Speech Recognition (ASR) systems have to deal with different kinds of variabilities with background noise being one of them. Besides feature representations that try to be immune to noise, e.g., RASTA-PLP [1], many methods have been proposed to compensate for the acoustic mismatch between training and testing data due to noise. Generally, these methods are either speech feature or model enhancement techniques, and some methods can be seen as hybrid approaches. Cepstral mean normalization, stereo piecewise linear compensation for environment (SPLICE) [2], and vector Taylor-series (VTS) expansion [3] are exemplary methods for feature enhancement methods. Generally, these methods try to remove the effects of noise from the feature vectors to reduce the mismatch between training and testing data. Parallel model combination (PMC) [4] is one example for the group of model enhancement techniques, where the parameters of the clean acoustic models are adapted such that they approximate the model parameters of training with corrupted speech. An advantage of feature enhancement methods are the smaller computational costs compared to the model adaptation techniques. A feature enhancement method that is based on maximizing the sharpness of the power distribution and on power flooring was recently proposed [5] and yields so-called power-normalized cepstral coefficients (PNCC). It was shown that the PNCC approach outperforms other common methods like PLP and VTS.

The size of speakers is another variability that generally leads to a mismatch between training and test data. More precisely, the vocal-tract length (VTL) is a parameter that can relate these differences between the speakers to each other. Common approaches like vocal-tract length normalization (VTLN) or maximum-likelihood linear regression (MLLR) are applied after the feature extraction stage to count for the distorting effects due to different VTLs. Similar to the feature-based noise-robustness methods as described above, there also exist methods that directly try to extract vocal-tract length invariant features. Generally, invariant feature extraction methods compute parametric representations from speech signals that are invariant to certain transformations. ASR systems may benefit from invariant features in several aspects: For example, the performance of speaker-independent ASR systems without any adaptation methods due to limited hardware resources may be increased. In case of ASR systems that already use speaker-adaptation methods, it has been shown that the additional use of invariant feature extraction may further increase the accuracy of those systems; one example of such an extraction method was proposed as invariant-integration features (IIFs) [6]. IIFs were originally designed for increasing the ASR systems' robustness to the effects of VTL changes which naturally occur between individual speakers.

In practice, ASR systems must be robust to several types of variability at the same time. As will be described in the next section, a combination of the processing chain of PNCCs with the one of IIFs is possible and promises to yield features that are not only robust to noise, but also to the effects of VTL changes. This work investigates on the one hand the noise-robustness of the originally presented IIFs and presents experimental results for different noisy conditions. These are compared to the results of mel frequency cepstral coefficients (MFCCs), PLPs, and PNCCs. On the other hand, it is shown that the accuracies of the IIFs under noisy conditions, as well as in mismatching training-testing conditions with respect to the mean VTL under noisy conditions can be improved when the PNCC-principles are combined with the original IIF computation.

In the next section we give a brief summary of the computation of IIFs and the idea of "power-bias subtraction". Also, the combined processing chain is motivated by observations made with estimated probability density functions (pdfs) of individual features. The third section explains the experimental setting and the results. Conclusions are given in Section 4.

2. Review of Invariant-Integration Features and Power Normalization

The processing chain for the computation of IIFs is shown in Figure 1, where the originally presented computation follows path (a). For an efficient computation of the time-frequency (TF) representation, a gammatone filter bank is used which is based on the weighting of the magnitude of a short-time Fourier transformation (STFT) with gammatone filter shaped coefficients. The frequency centers of the filters are linearly spaced on the ERB scale. Empirically determined, a power-law compression with an exponent of 0.1 is applied on the spectral values to resemble the nonlinear compression found in the human auditory system. Interestingly, the use of a gammatone filter bank together with the

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

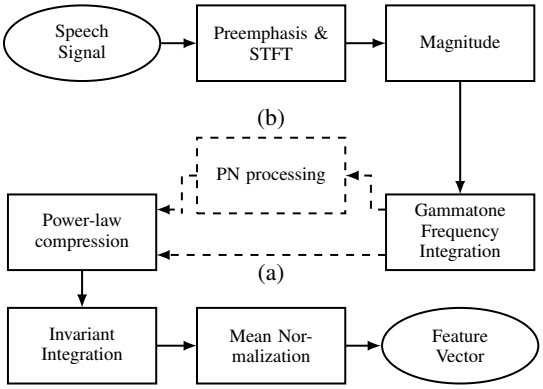

Figure 1: Processing chain for the computation of invariant integration features. (a) originally proposed computation [6]. (b) The combined computation as investigated in this work.

same power-law nonlinearity was also proposed in the original work on PNCCs [7], where this approach also showed superior properties in comparison to a standard mel filter bank with triangular shaped filter weights.

After the application of the nonlinearity, an invariant integration according to [6] is performed. Invariant integration is a general approach for the construction of invariants for arbitrary transformation groups. Generally, its computation involves the integration of (possibly nonlinear) functions m over all possible transformed observations. For the group of discrete translation, which approximately occurs in ASR as effect of VTL changes (e.g. [8, 9]), it was shown in [10] that the use of monomials as m yields a complete transformation.

Practically, with a given TF representation $y_k(n)$, where n is a frame index and k is a subband index, the monomials are defined as [6]

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

where $\vec{k} \in \mathbb{N}^M$, $\vec{l} \in \mathbb{N}_0^M$, and $\vec{m} \in \mathbb{N}^M$ describe the used subbands, integer components, and temporal offsets, respectively, and w is a subband-index offset. Following the integral approach, a monomial is evaluated on several translated versions of each frame and their results are averaged over a window size $2W + 1$:

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

The final feature vector $\vec{A} \in \mathbb{R}^N$ is a composition of several of these averages,

$$\vec{A}(n) = (A_{\hat{m}_1}(n), A_{\hat{m}_2}(n), \ldots, A_{\hat{m}_N}(n)). \quad (3)$$

A component-wise mean subtraction is the last step of the IIF computation. The parameters of the IIFs where determined with an iterative feature selection method that is based on a linear classifier [11]. In [6] it was shown that already an IIF set with appropriately chosen monomials of order one yields accuracies that outperform MFCCs significantly in matching average VTL training-test conditions as well as in mismatching VTL training-test conditions.

The computation of PNCCs also involves in its first steps the computation of a TF representation with a gammatone filter bank, which is the same as for IIFs. Then, the PNCC approach uses a power-normalization based on the $95th$ percentile and a "power-bias subtraction" (PBS) [5], which aims to maximize the sharpness of the power distribution and thereby minimize the acoustic mismatch due to noise. After applying a power-law compression, the discrete cosine transform (DCT) and a subsequent mean subtraction are used as final feature extraction steps for the computation of PNCCs.

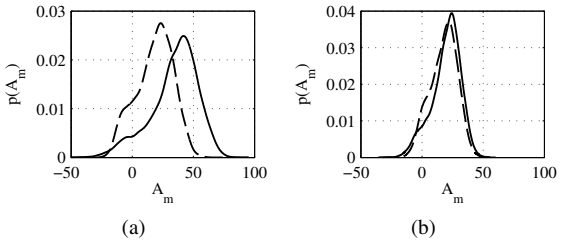

Figure 2: Estimated pdfs for exemplary IIF with clean speech (solid) and with noisy speech (dashed, SNR 10 dB) without PN-processing (a) and with PN-processing (b).

While both approaches are similar in their computation, IIFs and PNCCs have their emphasis on two different stages within the feature extraction process: While the PNCC approach concentrates on the enhancement of the noise-distorted TF representation, IIFs replace the DCT with an invariant-integration, such that shifts in the TF representation due to different VTLs have a minimal distorting effect in the feature space. Because these shifts are still present after the PN-processing, the two approaches can be combined with each other as depicted in Figure 1 by following path (b). This path involves an additional processing step compared to the original processing chain and is denoted as "power-normalization (PN) processing".

As an illustration of the effects of the combined processing, Figure 2 shows estimated pdfs of an original IIF for clean and noisy speech without and with PN-processing enabled. For both cases clean speech and noisy speech with an SNR of 10 dB were considered. It can be seen that the acoustic mismatch between clean and noisy speech features is generally smaller with the combined feature type than with the originally proposed IIFs. For quantitative measures, phoneme recognition experiments under different noise conditions and different training-testing scenarios with respect to the mean VTL were conducted. These experiments are described in the following.

To benefit from the theoretic advantages of IIFs, an appropriate selection of monomial parameters is crucial. It was shown in previous works of the authors [6] that a feature selection method based on the mean square error of a linear classifier yields features that outperform MFCCs when the TIMIT training set was used for selection. Within the experiments of the present work, no clear consistency between the relevance measure of the feature selection method and the resulting recognition accuracies were observed when the monomial parameters where selected on base of distorted TIMIT speech signals. Therefore, optimal parameter selection for noisy data will be further investigated in future work. For the remainder of this paper, we used parameters that proved to work well on clean data.

3. Experiments

Experiments were conducted on the TIMIT corpus with a sampling rate of 16 kHz to allow for the comparison to previous results. We used the NIST standard training set and the NIST complete test set which excludes the dialect (SA) sentences. The training set consists of 462 female and male speakers and con-

tains 3696 utterances. The test set contains 1344 utterances from 168 speakers with no overlap between training and test sets. To allow for an assessment of the performance of the feature types under mismatching training-test conditions with respect to the average VTL, two different scenarios were defined: The matching VTL scenario refers to the standard training and test sets. The mismatching VTL scenario uses only the male utterances from the training set for training and only the female utterances from the test set for testing. Phoneme recognition experiments with these two scenarios were conducted under different noise conditions. Clean and matched-condition training were considered in the experiments. The distorted speech signals were generated with the tool "FanT" as it was used for the AURORA corpus. All noise experiments were conducted with two different types of environmental noise, namely noise recorded at a station platform and noise recorded in a busy shopping mall[1]. In the following, the average accuracies obtained with the two noise types are presented.

The toolkit HTK was used for the training of the acoustic models as well as for decoding. A bigram language model based on the TIMIT training set was used in all experiments. Tied-state triphones with diagonal covariance modeling were used. The number of Gaussians in the mixtures of the individual models was chosen relative to the available size of the training data. While a maximum number of 16 Gaussians was used for MFCCs, RASTA-PLPs, and PNCCs, a maximum number of 8 Gaussians was used for IIFs. Following the standard procedure for TIMIT, the initial 61 phonetic labels were collapsed to a set of 48 labels. For testing, the phonetic labels were further collapsed to 39 labels. In case of PNCCs, the parameters as proposed in [5] were used. All feature vectors were concatenated with the log-energy and also with their first- and second-order time derivatives.

For the computation of the MFCCs, the standard HTK implementation was followed and 12 coefficients (with cepstral mean subtraction) were computed for each frame. For the computation of the PNCCs, the implementation from [5] was taken. For the comparison with the originally proposed IIFs, the IIF set consisting of 30 features that is based on a 110-band TF-representation from [6] was taken. These features are denoted as IIF_{Orig} in the following. In case of the IIFs, the 30-component feature vector was first reduced with a linear discriminant analysis (LDA) to 20 dimensions and then concatenated with the log-energy feature and the delta features. The LDA used the whole training dataset. Hence, the final feature dimensionality of the IIF-based system was 63. A maximum-likelihood linear transformation (MLLT) was computed to allow for diagonal covariance modeling. In all experiments, normalization and adaptation with VTLN and MLLR were applied for speaker-adaptive training and for testing.

3.1. Baseline results

The first part of the experiments presents baseline recognition accuracies and thereby compares the accuracies between the originally presented IIFs, the standard feature types MFCCs and RASTA-PLPs, as well as PNCCs under noisy conditions and for the matching and for the mismatching VTL scenarios. Figure 3 shows the results of these experiments.

It can be observed that for clean speech the IIFs perform best in both training-testing scenarios, and the accuracies resemble the ones presented in [6]. Looking at the performance of all feature types under noisy conditions, it can be seen that MFCCs yield the lowest accuracies compared to the other feature types. Besides IIFs, the PLPs show overall a superior noise-robustness to MFCCs, and PNCCs in turn a superior noise-robustness to PLPs. In terms of the lateral threshold shift, PNCC processing provides a maximum improvement in the matching and in the mismatching VTL scenario of about 7 dB, when compared to MFCCs. This holds for both scenarios and confirms the findings of [5], in this case on the TIMIT corpus. With respect to the IIFs, the experimental results show comparable accuracies to the PNCCs for an SNR of 20 dB for both scenarios. However, the performance of the IIFs aligns more and more with the one of PLPs when the SNR decreases.

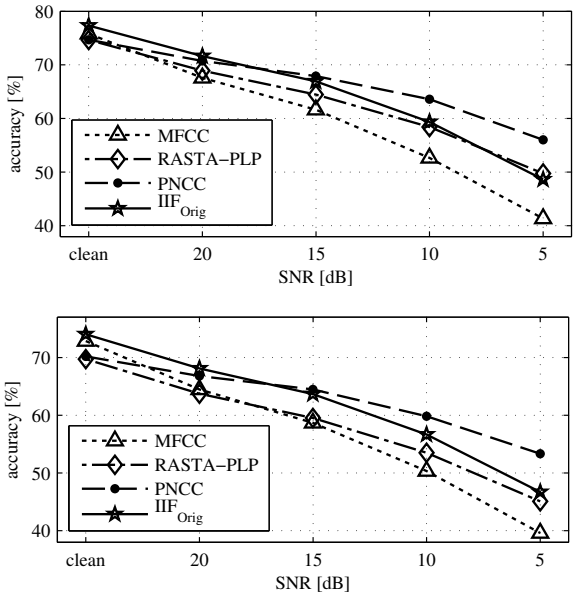

Figure 3: Baseline results for clean and noisy speech under matching (top) and mismatching (bottom) VTL training-test conditions (clean speech training).

3.2. PN-processed invariant-integration features

Originally, IIFs have specifically been designed for increasing the robustness of ASR systems under mismatching conditions with respect to the average VTL. Motivated by the observations made when PN processing is combined with the IIF computation (as described above and illustrated in Figure 2), we designed combined feature types in the second part of the experiments. Therefore, the default parameters of the power-bias subtraction algorithm were adjusted in preliminary experiments to better fit the characteristics of the used filter bank; with respect to the concepts described in [5], the power-flooring coefficient was set to 0.02, the medium-duration window factor was chosen as 3, and the weight-smoothing factor was set to 13. The results of these experiments are shown in Figure 4.

For clean speech, it can be observed that the accuracy of the combined feature type decreases by about one percentage point compared to the original IIFs (76.2% compared to 77.3%). However, these features still show better performance than PLPs and PNCCs in both scenarios. This decrease might originate from a still suboptimal parameter choice and will be part of further investigations. For noisy speech, the improved accuracies in both scenarios show that the noise robustness of the combined IIFs generally benefits from the additional PN-processing. For SNRs of 15 dB and above, the lateral threshold shifts to the PNCCs are about 5 dB and 7 dB for the matching and mismatching

[1]Noise signals are available for download at http://www.isip.uni-luebeck.de/downloads.

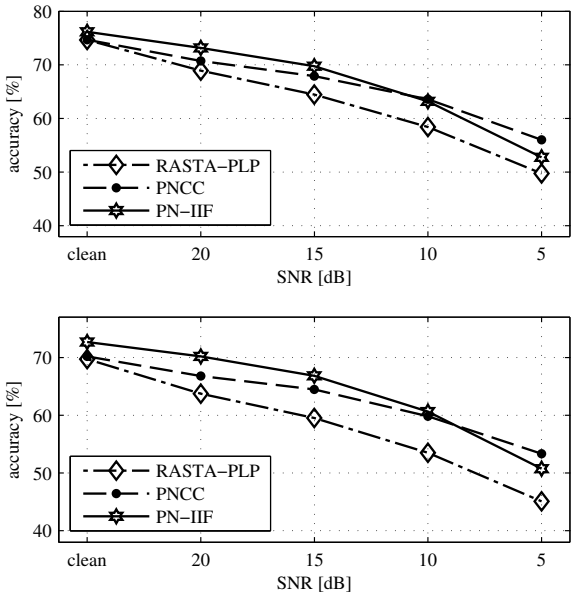

Figure 4: Results for enhanced features under matching (top) and mismatching (bottom) VTL training-test conditions (clean speech training).

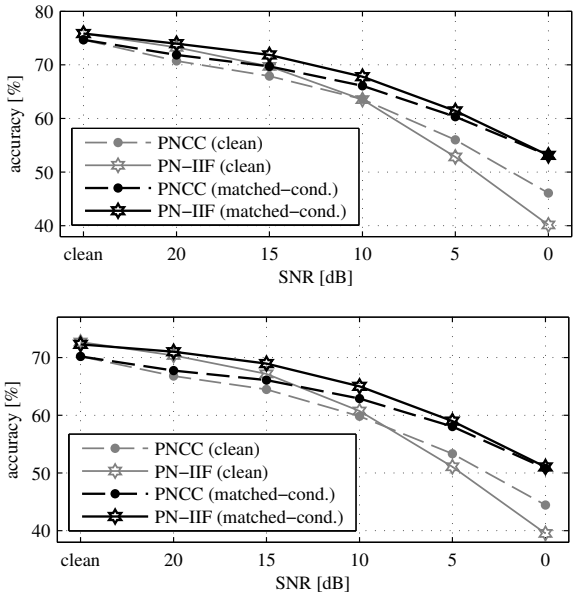

Figure 5: Results for enhanced features under matching (top) and mismatching (bottom) VTL training-test conditions (matched-condition training, black). For comparison, the gray lines indicate the accuracies for the clean-speech training case.

VTL scenarios, respectively. For distorted speech with an SNR of 10 dB, the combined features yield similar accuracies as the PNCCs, which is also a significant improvement. For an SNR of 5 dB, the PNCCs perform best.

Matched noise-condition training is a common approach to increase the noise-robustness of ASR systems. The idea is to decode noisy speech with acoustic models that were trained on data with equal noise-conditions. The last part of the experiments analyzed the performance of PNCCs and the features of the combined processing with this approach. The result of these experiments are shown in Figure 5. Besides the expected increase in accuracy under noisy conditions, it is interesting to see that the performance of the PN-processed IIFs did increase especially for SNRs of 5 and 0 dB and now perform equally well as PNCCs for low SNRs and significantly better for high SNRs. Overall, it can be observed that the combination of PN-processing and invariant integration leads to features that are both robust to noise and robust to varying VTLs.

4. Conclusions

In this paper we have presented new results that show that the invariant-integration features (IIF) without any feature-enhancement method for noisy conditions perform at least as good as RASTA-PLPs under noisy conditions. Furthermore, the processing chain for the computation of IIFs can be combined with the "power-bias subtraction" algorithm, such that the different benefits of both feature types are observable: Under clean conditions, the combined feature type shows better performance than MFCCs, PLPs, and PNCCs in matching as well as in mismatching VTL scenarios. When trained on clean speech only, but tested under noisy conditions, the combined processing also leads to the highest accuracies among the considered feature types for SNRs above 10 dB. For matched-condition training, the proposed features yield the best overall results. The computational cost for the combined processing is only slightly increased.

5. References

[1] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[2] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, Oct. 2000, pp. 806–809.

[3] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment independent speech recognition," in *Proc. Int. Conf. Audio, Speech, and Signal Processing*, vol. 2, Atlanta, USA, May 1996, pp. 733–736.

[4] M. J. F. Gales, "Model-based techniques for noise robust speech recognition," Ph.D. dissertation, Cambridge University, Cambridge, Sept. 1995.

[5] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *Proc. Int. Conf. Audio, Speech, and Signal Processing*, Dallas, USA, Mar. 2010, pp. 4574–4577.

[6] F. Müller and A. Mertins, "Contextual invariant-integration features for improved speaker-independent speech recognition," *Speech Communication*, vol. 53, no. 6, pp. 830 – 841, 2011.

[7] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. Interspeech 2009*, Brighton, UK, Sept. 2009, pp. 28–31.

[8] 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.

[9] 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.

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

[11] 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.