ON USING THE AUDITORY IMAGE MODEL AND INVARIANT-INTEGRATION FOR NOISE ROBUST AUTOMATIC SPEECH RECOGNITION

Florian Müller and Alfred Mertins

Institute for Signal Processing, University of Lübeck
Ratzeburger Allee 160, 23562 Lübeck, Germany

ABSTRACT

Commonly used feature extraction methods for automatic speech recognition (ASR) incorporate only rudimentary psychoacoustic findings. Several works showed that a physiologically closer auditory processing during the feature extraction stage can enhance the robustness of an ASR system in noisy environments. The “auditory image model” (AIM) is such a more sophisticated computational model. In this work we show how invariant integration can be applied in the feature space given by the AIM, and we analyze the performance of the resulting features under noisy conditions on the Aurora-2 task. Furthermore, we show that previously presented features based on power-normalization and invariant integration benefit from the AIM-based integration features when the feature vectors are combined with each other.

***Index Terms*—** Robust speech recognition, auditory processing, feature extraction, invariant integration

1. INTRODUCTION

For *automatic speech recognition* (ASR) systems, *mel frequency cepstral coefficients* (MFCC) [1] are well established speech signal representations and are used in many state-of-the-art ASR systems. The methods used for their extraction are mainly based on traditional engineering techniques and incorporate only rudimentary psychoacoustic findings. Besides speaker-extrinsic sources of variability like environmental noise or transmission-channel characteristics, there is the *vocal-tract length* (VTL) as one speaker-intrinsic variability that an ASR system has to deal with [2]. Different approaches that enhance the robustness of speaker-independent ASR systems are commonly part of practical ASR systems. While *vocal-tract length normalization* (VTLN) [3] normalizes the features extracted by the front-end, *maximum likelihood linear regression* (MLLR) [4] transforms the parameters of the acoustic models to better represent the characteristics of the individual speakers. Another group of methods tries to directly extract invariant features from the speech signal, using transforms that are invariant to the effects of VTL changes [5, 6, 7, 8].

With ASR systems still not reaching the recognition performance of human listeners, a more detailed imitation of the processes within the auditory system promises to further enhance the performance of ASR systems under certain conditions, e.g., noisy conditions [9]. The *auditory image model* (AIM) [10] tries to simulate the auditory processing of speech signals as they proceed up to the cochlear nuclei. As pointed out by [9] and described in more detail in the following, translation-invariant transforms fit well into the theoretic framework of the AIM and could be used to extract features with more robustness to VTL differences. We show how invariant-integration can be applied in the transform space of the AIM and analyze the performance of the resulting features under noisy conditions. Furthermore, we show that a recently presented combination of power normalization and invariant integration [11] benefits from the AIM-based IIFs when combining the feature types.

The article is organized as follows: The next section describes the general structure of the feature extraction and application of invariant integration within the signal space of the AIM. In Section 3 the experimental setup and results are presented. Discussion and conclusions are given in the last section.

2. THE AUDITORY IMAGE MODEL AND INVARIANT INTEGRATION

2.1. The Auditory Image Model

The AIM is a computational model that simulates human auditory processing. It aims to convert the speech signal into the perception that a human initially has before any semantic meaning is associated with the signal. The AIM is divided into several modules, which have either physical or psychoacoustic analogies. An illustration of the core structure of the AIM is shown within the gray shaded region in Fig. 1. A central element of the AIM is the strobed temporal integration stage whose functionality is comparable to that of a sparse autocorrelation function driven by the onsets of glottal pulses [12, 9]. The output of this stage is called *stabilized auditory image* (SAI) which yields a two-dimensional signal representation for every considered time step (see also Fig. 2). One dimension of the SAI is indexed by the subband number, and the second dimension by time intervals relative to the strobe-times index.

The effects of different VTLs within the SAI space have been analyzed in detail in [13], where it is shown that the $\{\text{scale, time}\}$-space of the SAI is scale covariant. This means that a change of the VTL leads to a shift along the subband axis, as well as to a scaling along the time-interval axis. The scaling is caused by the different lengths of the impulse responses of the filters.

The AIM has proved to yield beneficial auditory representations for various speech processing tasks: In [7] the AIM was used to extract low-dimensional features for speech recognition that are more robust to VTL changes than MFCC features on a synthetic speech corpus. In another work [12], sparse codes from the SAI were computed and used for sound retrieval and ranking. There exist different publicly available implementations of the AIM. Here, we used the Matlab version [14]. In the next section we describe how the scale covariant space of the SAI can be transformed to a scale/VTL invariant space with the help of invariant integration.

This work has been supported by the German Research Foundation under Grant No. ME1170/4-1.

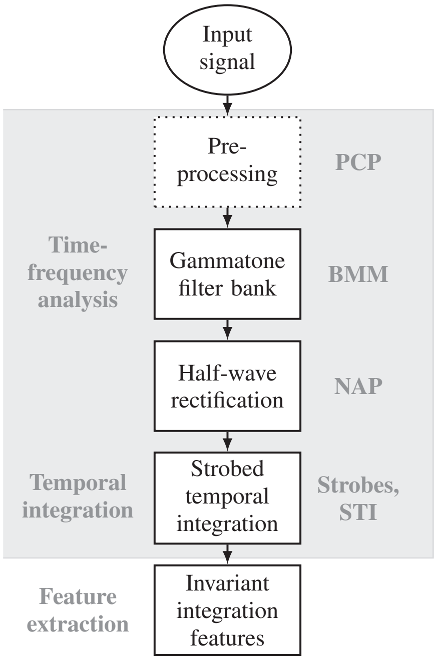

Fig. 1. Overview of the modular structure of the auditory image model (AIM, gray shaded area) [10], which consists of the pre-cochlear processing (PCP, not used in this work) module, basilar membrane motion (BMM) module, neural activity pattern (NAP) module, and the strobed temporal integration (STI) module, which makes use of a preceding strobe detection method.

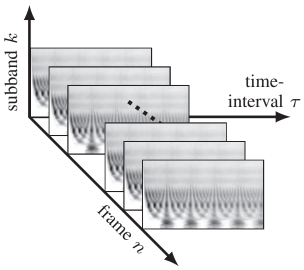

Fig. 2. Visualization of sequence of stabilized auditory images (SAIs) of a speech signal.

2.2. AIM-based Features for ASR with Invariant Integration

Cepstral coefficients are used in many state-of-the-art ASR systems due to their good performance and their efficient computation that involves only a few parameters. However, with respect to speaker independence, the cepstral analysis with an auditory filter bank has the disadvantage that different VTLs lead to translations along the subband axis, while the *discrete cosine transform* (DCT) is not translation invariant. Thus, the same phoneme uttered from two vocal tracts with different lengths do not yield the same point in the MFCC space. Generally speaking, a feature extraction method should only extract information that is necessary for separating the individual classes of interest and, at the same time, be invariant to the effects of other variabilities. Using the AIM for feature extraction for noise-robust ASR can further be motivated by the observations made in [9], were SAI-based features showed a larger noise robustness than MFCCs (while performing worse under clean-speech conditions).

Invariant integration is a constructive approach for computing separable features that are invariant to a designated group of transformations. For its application in speaker-independent ASR tasks, so-called *invariant-integration features* (IIFs) were described in detail in [8]. The key concept of the IIFs is their invariance to translations along the subband axis. Motivated by the observation that the SAI space is scale covariant, we propose a new definition for invariant-integration features based on the SAI space in this work. Therefore, one has to integrate over the induced transformation due to different VTLs in this space. In the following, we first give a short formal introduction to IIFs based on a standard mel or gammatone filter bank. With the then introduced terms, it is described afterwards how SAI-based IIFs can be computed.

We follow the notation as introduced in [8]: Let $v_k(n)$ denote the TF representation of a speech signal, where n is the time index, $1 \leq n \leq N$, and k is the subband index with $1 \leq k \leq K$. We define the vectors $\boldsymbol{k} = (k_1, k_2, \ldots, k_M)$ and $\boldsymbol{l} = (l_1, l_2, \ldots, l_M)$, containing element indices and integer exponents with $\boldsymbol{k} \in \mathbb{N}^M$ and $\boldsymbol{l} \in \mathbb{N}_0^M$, respectively. Furthermore, let $\boldsymbol{m} \in \mathbb{Z}^M$ be a vector containing temporal offsets. Now, we define a contextual monomial $\hat{m}$ with M components as

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

where $\gamma(\boldsymbol{l}) := \sum_{i=1}^{M} l_i$ is a normalizing term and is referred to as the "order of the monomial". Also, $w \in \mathbb{N}_0$ is a spectral offset parameter that is used for ease of notation in the following definitions. Now, a single IIF is defined as

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

with W determining the window size. An adequately chosen IIF set of size F,

$$\mathbf{A} := \{A_{\hat{m}_1}, A_{\hat{m}_2}, \ldots, A_{\hat{m}_F}\}, \tag{3}$$

yields features that, on the one hand, are invariant to the translations as (approximated) spectral effects due to VTL changes, and, on the other hand, allow for discriminating between the individual classes. For determining the parameters of a set of IIFs an iterative feature selection method is used that is based on a linear classifier [8].

To apply the invariant integration approach within the SAI space, one has to integrate over the induced transformation due to different VTLs within the SAI space. That is, the scaling effect between the subbands with different VTLs has to be considered. This relation can be described with the product of time interval and center frequency of the individual subbands being constant [15]: Let $\tilde{v}_k(n, \tau)$ denote the SAI value at time instance n, with k being the subband index, and τ being the time interval. Furthermore, let $\boldsymbol{c} = (c_1, c_2, \ldots, c_K)$ denote the center frequencies of the filters. Now, for a given subband index $i \in \mathbb{N}$ and a cycle number $p \in \mathbb{R}^+$,

$$\tau_i(p) := \frac{p}{c_i} \tag{4}$$

defines the time interval for each subband, which corresponds to the same cycle for all impulse responses. Now, the SAIs of the same utterance from two speakers A and B with different VTLs are related by

$$\tilde{v}_i^A(n, \tau_i(p)) = \tilde{v}_{i+\alpha_T}^B(n, \tau_{i+\alpha_T}(p)), \tag{5}$$

where α_T is proportional to the ratio between the VTLs of A and B. Thus, a change of VTL leads to a shift of the formants along ridges which pass through the same cycles of the impulse responses of all subbands. In [13, 15] the representation of the SAI in this $\{$scale-cycle$\}$-space is called *size-shape image* (SSI). The SSI space is scale-shift covariant, which means that the effects due to different VTLs appear solely as translations along the subband axis.

Now, let $\boldsymbol{p} = (p_1, p_2, \ldots, p_M)$ contain cycle numbers. We define a monomial $\tilde{m}$ on base of the SAI space as

$$\tilde{m}(n; w, \boldsymbol{k}, \boldsymbol{l}, \boldsymbol{m}, \boldsymbol{p}) := \left[\prod_{i=1}^{M} \tilde{v}_{k_i+w}^{l_i}(n+m_i, \tau_{k_i+w}(p_i))\right]^{1/\gamma(\boldsymbol{l})} . \tag{6}$$

With the definition from Eq. (6), a feature component $A_{\tilde{m}}(n)$ is then computed as defined in Eq. (2). The features based on the SAI will be referred to as AIM-IIFs in the following. We have used linear interpolation in this work to compute $\tilde{v}_i(n, \tau_i(p))$.

The idea of applying an invariant transform on the SAI was also part of [15], were an adapted form of the Mellin transform was used to compute a VTL-invariant representation of speech signals: The Mellin image [15] is essentially the magnitude of the Fourier transform of the corresponding SSI vector and, thus, is also translation invariant. However, compared to the approach proposed in this work, the Mellin image has at least two disadvantages: Though invariant to translations, the magnitude of the Fourier transform is also invariant to additional transformations like mirroring. Also, the data rate of the Mellin image is as high as the one of the SAI and, thus, would need to be reduced prior to be fed into an ASR system. A benefit of the IIFs is that only selected segments of constant cycle numbers need to be considered, so that a transformation of the complete SAI into an SSI is not necessary. Furthermore, the extraction of AIM-IIFs also leads to a significant reduction of the data rate, which is comparable to that of cepstral features.

3. EXPERIMENTS

3.1. Experimental Setup and Baselines

Experiments have been conducted on the Aurora-2 task. We used the standard training and test sets as they are published together with the corpus data. These include utterances with *signal-to-noise ratios* (SNR) of 20, 15, 10, 5,0, and -5 dB. Both clean speech and multi-condition training were considered. Average accuracies of all three test sets are shown in the following. Throughout the experiments, the same HTK-based back-end was used. Whole-word left-to-right models with 11 to 17 states depending on the average utterance lengths of the digits were used. Four Gaussians were used in the mixtures of the individual states, and the covariance matrices were constrained to be of diagonal form. All features were extracted with a frame rate of 100 Hz. First- and second-order time-derivative approximations were appended to all feature vectors. In case of integration features, a *linear discriminant analysis* (LDA) with a target dimensionality of 55 followed by a *maximum-likelihood linear transform* (MLLT) [16] was computed to reduce the feature vector dimensionality. The target dimensionality of 55 was empirically chosen in preliminary experiments.

Baseline accuracies were generated with MFCCs using the standard setup of HTK together with cepstral mean subtraction and also with *power-normalized cepstral coefficients* (PNCC) [17] as second feature type. PNCCs are cepstral coefficient-based features that can efficiently be computed and have recently shown a comparable noise-robustness in comparison to feature enhancement methods like Vector-Taylor series expansion or SPLICE. The results of these two feature types for both training modes are shown in Table 1.

Table 1. Baseline accuracies [%] for MFCCs and PNCCs for clean and multi-condition training on Aurora-2

SNR	**clean**		**multi-condition**	
(dB)	**MFCC**	**PNCC**	**MFCC**	**PNCC**
∞	98.6	98.6	98.4	98.0
20	96.8	97.7	97.9	97.7
15	93.0	95.7	96.8	97.0
10	78.1	90.1	93.9	95.1
5	51.2	75.7	85.2	88.3
0	26.3	49.1	65.1	72.1
-5	12.2	21.9	31.1	40.2
Avg.	65.2	75.5	81.2	84.1

Table 2. Accuracies [%] for IIFs, PN-IIFs, and AIM-IIFs for clean and multi-condition training on Aurora-2

SNR	**clean**			**multi-condition**		
(dB)	**IIF**	$\mathbf{IIF_{PN}}$	$\mathbf{IIF_{AIM}}$	**IIF**	$\mathbf{IIF_{PN}}$	$\mathbf{IIF_{AIM}}$
∞	99.2	98.8	98.5	98.7	98.9	98.5
20	97.8	97.8	97.4	98.3	98.5	98.2
15	95.4	96.0	95.2	97.4	97.9	97.6
10	89.5	91.0	88.0	95.6	96.0	95.7
5	72.8	77.7	70.1	89.2	90.3	88.9
0	41.0	50.4	40.5	68.1	72.9	67.4
-5	14.0	22.1	17.3	31.0	39.1	29.6
Avg.	72.8	76.2	72.4	82.6	84.8	82.3

The results clearly show the advantage of PNCCs compared to MFCCs under noisy conditions. While performing similar in case of clean speech, the PNCC-based ASR system achieves accuracies that are increasingly better in terms of accuracy the lower the SNR becomes. This holds for clean speech training as well as for multi-condition training and supports the results from [17]. To allow for a comparison with integration-based features, we show accuracies of IIFs [8] and PN-IIFs [11] in Table 2. PN-IIFs combine the methods for increasing the noise robustness of the PNCCs with invariant integration, which further increases the robustness to the effects of VTL differences. Table 2 also shows the results for the AIM-IIFs. For the feature selection, the same method as for the “standard” IIFs was used. We selected sets of 30 AIM-IIFs with the constraint of using only at most a cycle number of three. This constraint was used, because the glottal pulse rate imposes an upper limit for the time-interval before resonance and pulse information are superimposed in this space [13, 15].

The IIF-based ASR system achieves accuracies that are higher than the MFCC-based system for all SNRs and for both training conditions. In comparison to PNCCs, however, they do not perform as well, which was also observed in [11] and motivated PN-IIFs. PN-IIFs perform better than IIFs under all SNRs and better than PNCCs for SNRs down to 5 dB for both training modes. A reason for the abrupt decrease of accuracy for lower SNRs in case of the PN-IIFs might be the fact that the reduction matrix, which is estimated with LDA on the training data, does not generalize well for noise scenarios with very low SNRs. Also, the parameters chosen for the power-normalization feature-enhancement stage might still not be optimal. A comparable performance of these feature types was also observed on artificially distorted TIMIT speech signals as presented in [11].

Table 3. Accuracies [%] for AIM-IIFs combined with originally proposed IIFs, PNCCs, and with PN-IIFs for clean and multi-condition training on Aurora-2

	clean			**multi-condition**		
SNR	**IIF_{AIM}**			**IIF_{AIM}**		
(dB)	**+IIF**	**+PNCC**	**$+\text{IIF}_{\text{PN}}$**	**+IIF**	**+PNCC**	**$+\text{IIF}_{\text{PN}}$**
∞	99.2	99.0	99.0	98.8	98.7	98.9
20	97.9	97.8	98.2	98.2	98.2	98.5
15	96.0	96.5	96.7	97.4	97.6	98.0
10	90.9	92.3	92.6	95.5	96.2	96.4
5	77.3	81.1	81.3	89.8	91.3	91.6
0	48.4	54.7	57.0	71.9	75.5	76.4
-5	17.9	21.7	26.3	33.9	39.8	43.4
Avg.	75.4	77.6	**78.7**	83.7	85.3	**86.2**

3.2. Results for AIM-IIFs and Feature Combinations

The AIM-IIF-based ASR system shows a comparable performance to the IIF-based system for both training conditions in Table 1. In [9] it was observed that SAI-based features yield a higher robustness under noisy conditions compared to MFCCs, while performing worse under clean conditions. This cannot be observed for the results in Table 2. However, due to the different way of processing the speech signal in comparison to the standard filter banks, we assume that SAI-based signal representations are prone to different kinds of errors. Thus, we investigated if a combination of AIM-IIF vectors together with IIF, PNCC, or PN-IIF feature vectors yields an enhancement in accuracy. In contrast to ROVER, this approach has the advantage that only a single ASR system is used. Again, we used LDA and MLLT to reduce the resulting feature dimension down to 55 and to decorrelate the features. The results of these experiments are shown in Table 3. It can be observed that the concatenation of all three feature types with AIM-IIFs generally increases the accuracy. Combining AIM-IIFs and PN-IIFs into one feature vector yields the highest accuracies within the experiments of this work and gives average accuracies (over all SNRs) of 78.7% and 86.2% for clean speech and multi-condition training, respectively.

4. CONCLUSIONS AND OUTLOOK

The AIM is a computational model of the human auditory processing pathway, which represents a speech signal at every time instance within a two-dimensional space that is scale covariant. Motivated by other works that used the SAI as basis for extracting features for ASR, we showed how the concept of invariant-integration can be applied within the SAI-space. We conducted ASR experiments under different noise conditions on the Aurora-2 task. When used as sole features, the SAI-based IIFs (“AIM-IIFs”) performed equally well compared to other state-of-the-art feature types. The significant increase in accuracy when combined with another integration-based feature type, however, suggests that the AIM-IIFs yield complementary information compared to the other integration feature types.

The parameters of the AIM offer a high degree of freedom. Though empirically determined in preliminary experiments, the used parameters for the AIM might still have room for optimization. Also, future work will investigate different reduction methods, which might perform better than LDA under noisy conditions. Due to the artificially distorted utterances of Aurora-2, this work can only be seen as a proof of concept. An evaluation on more naturally distorted speech would also take other effects, e.g., the Lombard effect, into account and is part of future work. A tool for the computation of IIFs can be downloaded at http://www.isip.uni-luebeck.de/downloads.

5. REFERENCES

[1] 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.

[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] 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.

[5] S. Umesh, Leon Cohen, Nenad Marinovic, and Douglas J. Nelson, “Scale transform in speech analysis,” *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 1, pp. 40–45, Jan. 1999.

[6] Jan Rademacher, Matthias Wächter, and Alfred Mertins, “Improved warping-invariant features for automatic speech recognition,” in *Proc. Int. Conf. Spoken Language Processing (Interspeech 2006 - ICSLP)*, Pittsburgh, PA, USA, Sept. 2006, pp. 1499–1502.

[7] Jessica J. Monaghan, Christian Feldbauer, Tom C. Walters, and Roy 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.

[8] Florian Müller and Alfred Mertins, “Contextual invariant-integration features for improved speaker-independent speech recognition,” *Speech Communication*, vol. 53, no. 6, pp. 830 – 841, 2011.

[9] Thomas C. Walters, *Auditory-Based Processing of Communication Sounds*, Ph.D. thesis, University of Cambridge, 2011.

[10] 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., Pergamon, Oxford, 1992, vol. 83, pp. 429–446.

[11] Florian Müller and Alfred Mertins, “Noise robust speaker-independent speech recognition with invariant-integration features using power-bias subtraction,” in *Proc. Interspeech-2011*, Florence, Italy, Aug. 2011, pp. 1677–1680.

[12] Richard F. Lyon, Martin Rehn, Samy Bengio, Thomas C. Walters, and Gal Chechik, “Sound retrieval and ranking using sparse auditory representations,” *Neural Computation*, vol. 22, pp. 2390–2416, 2010.

[13] R.D. Patterson, R. van Dinther, and T. Irino, “The robustness of bio-acoustic communication and the role of normalization,” in *Proc. Int. Congress on Acoustics*, Madrid, Sept. 2007, pp. ppa–07–011.

[14] S. Bleeck, T. Ives, and R. D. Patterson, “Aim-mat: the auditory image model in MATLAB,” *Acta Acustica United with Acustica*, vol. 90, pp. 781–788, 2004.

[15] T. Irino and R. Patterson, “Segregating information about the size and the shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform,” *Speech Communication*, vol. 36, no. 3, pp. 181–203, Mar. 2002.

[16] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, “Maximum likelihood discriminant feature spaces,” in *Proc. Int. Conf. Audio, Speech, and Signal Processing*, Jun. 2000, pp. 1129–1132.

[17] Chanwoo Kim and Richard 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.