applied system innovation

MDPI

Article

Employing Robust Principal Component Analysis for Noise-Robust Speech Feature Extraction in Automatic Speech Recognition with the Structure of a Deep Neural Network

Jeih-weih Hung *, Jung-Shan Lin and Po-Jen Wu

Department of Electrical Engineering, National Chi Nan University, 545, Nantou County, Taiwan; jslin@ncnu.edu.tw (J.-S.L.); tito19931221@gmail.com (P.-J.W.)

* Correspondence: jwhung@ncnu.edu.tw

Received: 6 April 2018; Accepted: 12 August 2018; Published: 15 August 2018

Abstract: In recent decades, researchers have been focused on developing noise-robust methods in order to compensate for noise effects in automatic speech recognition (ASR) systems and enhance their performance. In this paper, we propose a feature-based noise-robust method that employs a novel data analysis technique—robust principal component analysis (RPCA). In the proposed scenario, RPCA is employed to process a noise-corrupted speech feature matrix, and the obtained sparse partition is shown to reveal speech-dominant characteristics. One apparent advantage of using RPCA for enhancing noise robustness is that no prior knowledge about the noise is required. The proposed RPCA-based method is evaluated with the Aurora-4 database and a task using a state-of-the-art deep neural network (DNN) architecture as the acoustic models. The evaluation results indicate that the newly proposed method can provide the original speech feature with significant recognition accuracy improvement, and can be cascaded with mean normalization (MN), mean and variance normalization (MVN), and relative spectral (RASTA)—three well-known and widely used feature robustness algorithms—to achieve better performance compared with the individual component method.

Keywords: robust principal component analysis; noise robustness; filter-bank features; mel-frequency cepstral coefficients; deep neural network

1. Introduction

Automatic speech recognition (ASR) applications have been widely seen in our daily life, and some examples include voice-based command controls of a robot, speech recognition using mobile devices and speech-related web search. However, an ASR system usually degrades significantly in performance when it is applied to an environment with interferences such as additive noise and channel distortion. In recent decades, researchers have been focused on developing noise-robust methods in order to compensate for the aforementioned interference effects and enhance the ASR performance, and these methods can be roughly classified into two fields: feature-based and model-based.

Generally speaking, most of the feature-based methods are developed to enhance the noise-robust ability of existing and widely used speech features like perceptual linear prediction coefficients (PLP) [1], mel-frequency cepstral coefficients (MFCC) [2], logarithmic mel-filter-bank coefficients (FBANK) [2], and Gammatone frequency cepstral coefficients [3], and some of them act at the intermediate stage of the creating process of speech features. For example, spectral subtraction (SS) [4,5], Wiener filtering [6], and MMSE-based short-time spectral amplitude estimation [7,8] are exemplary methods that process the frame-wise acoustic spectra, which are then converted to final

features for speech recognition. Second, a variety of statistical moment normalization methods are developed and conducted on the intermediate and final stages of creating speech features to give significant improvement in recognition accuracy under noise-corrupted situations, such as mean normalization (MN) [9], mean and variance normalization (MVN) [10], and histogram normalization (HEQ) [11,12], to name but a few.

Furthermore, because the aforementioned statistical moments are directly evaluated by the temporal series of speech features, these moment normalization methods implicitly enhance speech features with regard to temporal characteristics. By contrast, another direction of feature enhancement directly and explicitly processes the temporal series of speech features, and the respective methods include, but are not limited to, RASTA [13], temporal structure normalization (TSN) [14], and MVN plus ARMA filtering (MVA) [15]. Additionally, the methods of spectral histogram equalization (SHE) [16], modulation spectrum replacement/filtering (MSR/MSF) [17], and nonnegative matrix factorization (NMF)-based modulation spectrum enhancement [18] directly modify modulation spectra, which are specifically referred to as the Fourier transform of the feature time sequence.

On the other hand, the model-based methods attempt to adapt the existing acoustic models with noise information to make them more suitable for an application environment. According to [19], these methods can be further split into two schools: general adaptation and noise-specific compensation. The general-adaptation methods use a generic transformation to convert acoustic model parameters, and some representative methods include maximum-likelihood linear regression (MLLR) [20], maximum likelihood linear transform (MLLT) [21], minimum classification error–based linear regression (MCELR) [22,23] and discriminative mapping transform [24]. By contrast, the noise-specific compensation methods update acoustic model parameters by explicitly adopting the characteristics of the noise present in an application environment, and they include parallel model combination (PMC) [25] and model-based vector Taylor series (VTS) [26]. Interested readers are referred to [19,27] for a comprehensive coverage of recent noise-robust techniques for automatic speech recognition.

In this paper, we propose a feature-based method that uses the technique of robust principal component analysis (RPCA) [28,29] aiming to extract noise-robust speech features. RPCA is a novel data analysis method and has been widely used for speech enhancement and robust speech representation algorithms, among which some well-known examples are briefly described here. In [30], RPCA is applied to the spectrogram of speech signals, and the resulting sparse component is shown to contain less noise and thus be noise-robust. Another speech enhancement method proposed in [31] first decomposes a speech signal into sub-bands via a wavelet transform and then uses RPCA to extract the low-rank component of the matrix created by the overlapped frames of each sub-band signal, and the final output is the inverse wavelet transform of low-rank sub-band signals. The method in [32] integrates RPCA and exemplar-based sparse representation in an SNR-dependent manner to process the spectrogram of a noise-corrupted signal, i.e., to use RPCA in a low-SNR case and use exemplar-based sparse representation in a high-SNR case. In [33,34], RPCA is also used to decompose the spectrogram of a noise-corrupted signal, while the respective sparse component is further constrained to be a nonnegative weighted sum of pre-learned basis spectra. Briefly speaking, the algorithms in [30,32–34] apply RPCA in the spectro-temporal (spectrogramic) domain of speech signals, and in [31], RPCA is operated in the time domain of speech signals within each sub-band produced by a wavelet transform.

The newly proposed method differs from the aforementioned RPCA-wise algorithms mainly in that it employs RPCA in the temporal series of FBANK/MFCC speech features, which are directly used in automatic speech recognition. Notably FBANK/MFCC features are a nonlinear transform of the spectrogram of a time-domain speech signal due to the logarithmic operation. In the proposed scenario, each signal in the training and testing sets for an ASR system is converted to FBANK/MFCC features. Then, the feature time sequence, expressed in a matrix form, is decomposed by RPCA to produce a sparse matrix and a low-rank matrix. Finally, the obtained sparse matrix is treated as the new features for the subsequent training or testing. Compared with the matrix that contains

the original features, the sparse matrix is shown to highlight the relatively fast-varying component, which very probably corresponds to the speech-dominant elements and benefits speech recognition. In comparison, the associated low-rank matrix reveals more static characteristics that are likely related to the embedded noise. We evaluate the proposed RPCA-based novel feature extraction method on the Aurora-4 benchmark task [35], which consists a medium-to-large vocabulary database and a recognition task based on the Wall Street Journal (WSJ) corpus [36]. In addition, state-of-the-art deep neural network (DNN) architecture is used for acoustic modeling in the experiments. The evaluation results show that the proposed RPCA-based method can provide the original feature with significant recognition accuracy improvement, and the achieved relative word error rate reduction can be as high as 43%. We also show that this new method can be additive to the prevalent mean normalization (MN) and relative spectral (RASTA) methods to further improve the recognition performance. As a result, these evaluation results indicate that the newly proposed method is quite promising to enhance the ASR and can broaden the corresponding applications in real environments.

This paper is organized as follows. In Section 2, we briefly introduce the RPCA algorithm. Section 3 includes the proposed RPCA-based feature extraction method. The experimental setup is described in Section 4, and the experimental results as well as the corresponding discussions and analyses are given in Section 5. Finally, Section 6 contains concluding remarks and suggestions for future work.

2. Robust Principal Component Analysis (RPCA)

Robust principal component analysis (RPCA), as the name suggests, is a modification of the well-known data analysis method, principal component analysis (PCA). However, RPCA is shown to perform well for noise-corrupted data by reducing the deteriorating effect of outliers compared with PCA. The central idea of RPCA is to decompose a data matrix V into another two matrices, L and S, as follows:

$$V = L + S, \tag{1}$$

where L is a low-rank matrix and S is a sparse matrix with most of its entries being zero. One of the most widely used methods to achieve the above decomposition is the Principal Component Pursuit (PCP) method [28], which solves the constrained optimization problem below.

$$\min(\|L\|_* + \lambda\|S\|_1) \text{ subject to } L + S = V \tag{2}$$

where $\|L\|_*$ is the sum of singular values of matrix V, $\|S\|_1$ is the sum of all entries of matrix S, and λ is a weighting factor set to $\frac{1}{\sqrt{n}}$, n being the dimension of matrix V.

According to [28], the two matrices L and S in Equation (2) can be determined accurately without any prior knowledge of them. For more details about the PCP method, one can refer to the literatures [28,29].

3. Proposed Method

In the presented method, the RPCA algorithm is applied to the matrix organized by the speech feature temporal sequence in order to extract the embedded noise-robust component. The procedure for this method is split into the following two steps:

Step 1: Create the baseline features, FBANK and MFCC:

Any time-domain utterance $\{x[\ell]\}$ in the training and test sets is first passed through a high-pass pre-emphasis filter, and the operations of framing and windowing are performed in turn. Then, each windowed frame signal is converted to the acoustic frequency domain via short-time Fourier transform (STFT) to create the corresponding acoustic spectrum. Next, each frame-wise acoustic spectrum is converted to the FBANK or MFCC features. The magnitude of the acoustic spectrum associated with each frame is weighted by a Mel-frequency filter bank and then processed with the

logarithmic operation to produce the FBANK features. Moreover, the MFCC features are derived after the application of the discrete cosine transform (DCT) to FBANK.

Step 2: Use RPCA to extract the sparse part of the FBANK/MFCC matrix

Let $\{\mathbf{v}_m; 0 \leq m \leq M-1\}$ denote the frame time series of FBANK or MFCC vectors for the time-domain signal $x[\ell]$ obtained from Step 1, where m is the frame index and M is the total number of frames. Then a matrix V is created by assigning the m^{th} column vector of V to be $\mathbf{v}_m$. Therefore, the matrix V is termed as the feature matrix of the time-domain signal $x[\ell]$.

Next, the RPCA algorithm stated in Section II is used to decompose the feature matrix V,

$$V = V_L + V_S, \tag{3}$$

where V_L and V_S denote the low-rank and sparse component matrices of V. Finally, we discard the low-rank part V_L while the sparse part V_S is preserved and treated as the new feature matrix for the subsequent processing in training and testing.

Because the main idea of the aforementioned method is to extract the sparse component of the feature matrix via RPCA, we will use the notation "RPCA-SPC" to denote this new method hereafter for the ease of discussion.

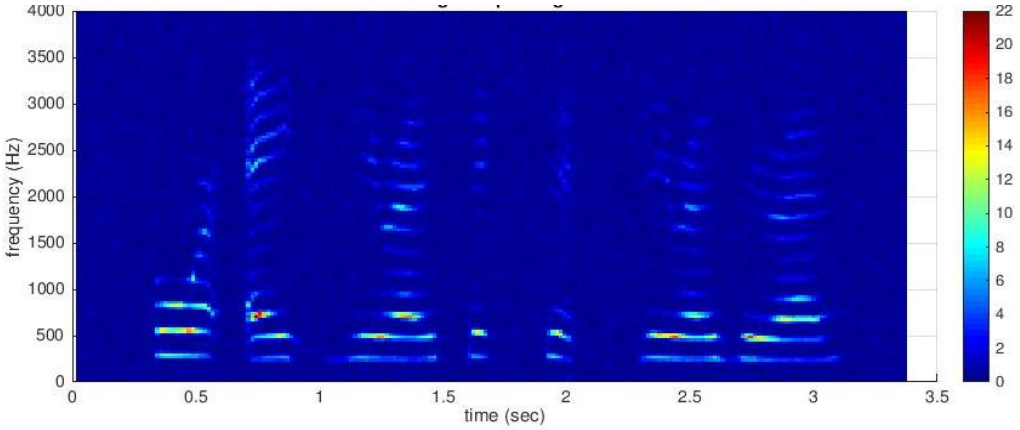
The primary idea of RPCA-SPC is as follows: For a noise-corrupted utterance and the corresponding speech feature sequence, the embedded noise part often varies more slowly with time relative to the clean-speech part. In other words, clean speech is likely to be more non-stationary than noise. This phenomenon can be easily observed when analyzing the spectrogram of an utterance. The spectral structure of pure noise is usually fixed or slow-varying, while the speech component changes quickly with time. Such an assumption implies that the noise part appears to be of low-rank, while the clean-speech part is sparse. Therefore, extracting the sparse component of the speech feature matrix tends to enhance the speech and alleviate noise.

Here we provide two examples to reveal that RPCA tends to highlight the clean-speech part in the noise-corrupted data:

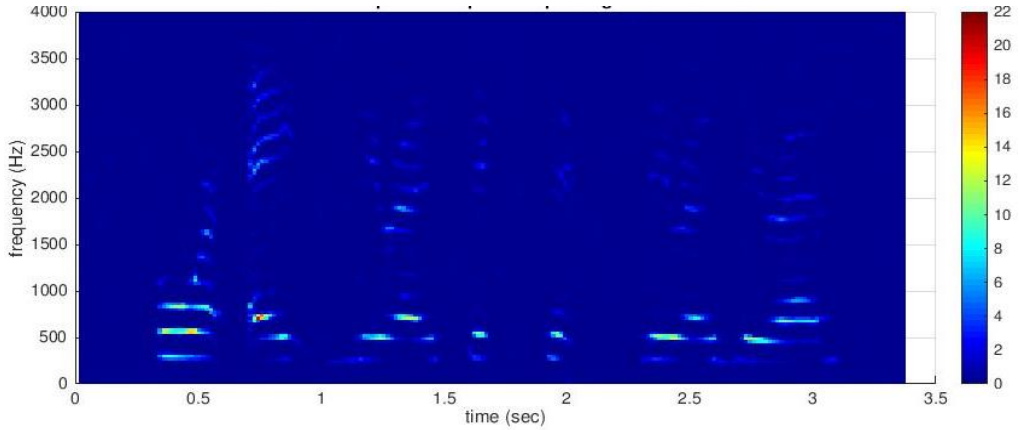
First, Figure 1a depicts the spectrogram of a noisy utterance, and Figure 1b,c depicts the RPCA-derived sparse and low-rank partitions of the spectrogram, respectively, shown in Figure 1a. From these figures, it is obvious that the sparse part contains rich speech clues, while the low-rank counterpart corresponds to relatively less speech information and more about noise.

Next, Figures 2a–c and 3a–c, respectively, show the time series of the sixth and eighth MFCC features of a noisy utterance as well as the corresponding sparse and low-rank components. Likewise, Figures 4a–c and 5a–c correspond to the original, sparse, and low-rank versions of the sixth and eighth FBANK coefficients of a noisy utterance. From these figures, it is clearly observed that the sparse component appears close to the original feature stream and shows synchronicity along the time axis to some extent, while the low-rank component behaves like irrelevant noise.

(**a**) **original spectrogram**



(**b**) **sparse-component spectrogram**



(**c**) **low-rank component spectrogram**

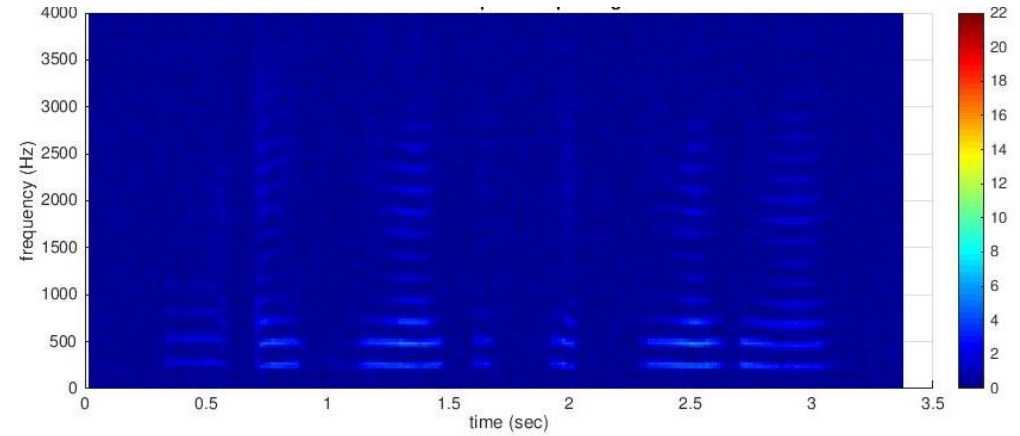


Figure 1. (**a**) The spectrogram of a noise-corrupted utterance which corresponds to an English digit string "four-eight-zero-six-six-zero-zero" and contains Gaussian random noise at an SNR of 10 dB. (**b**) The sparse part of (**a**) derived from RPCA. (**c**) The low-rank part of (**a**) derived from RPCA.

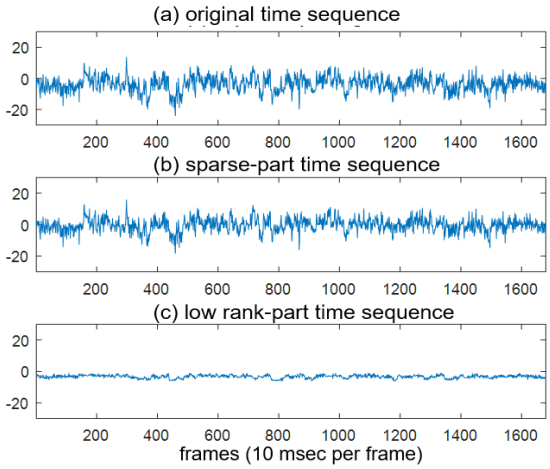


Figure 2. (**a**) The time series of the sixth mel-frequency cepstral coefficients (MFCC) of a noisy utterance; (**b**) The sparse part of (**a**) derived from RPCA; (**c**) The low-rank part of (**a**) derived from RPCA. Note: The RPCA is applied to the whole feature matrix of the utterance, and it holds for the cases in Figures 3–5.

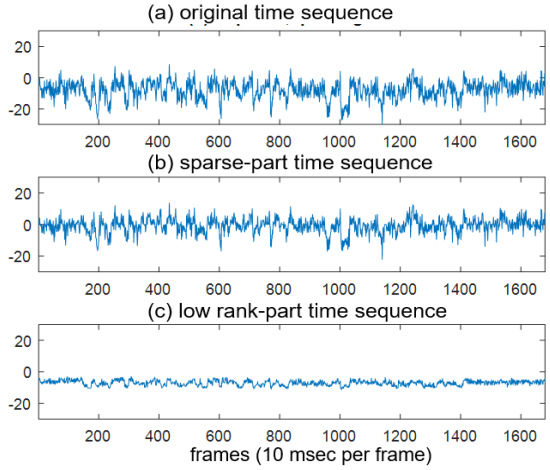


Figure 3. (**a**) The time series of the eighth mel-frequency cepstral coefficients (MFCC) of a noisy utterance; (**b**) The sparse part of (**a**) derived from RPCA; (**c**) The low-rank part of (**a**) derived from RPCA.

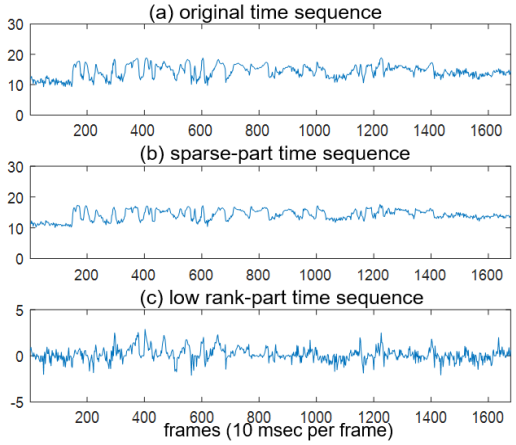


Figure 4. (**a**) The time series of the 6th logarithmic mel-filter-bank coefficients (FBANK) of a noisy utterance; (**b**) The sparse part of (**a**) derived from RPCA; (**c**) The low-rank part of (**a**) derived from RPCA.

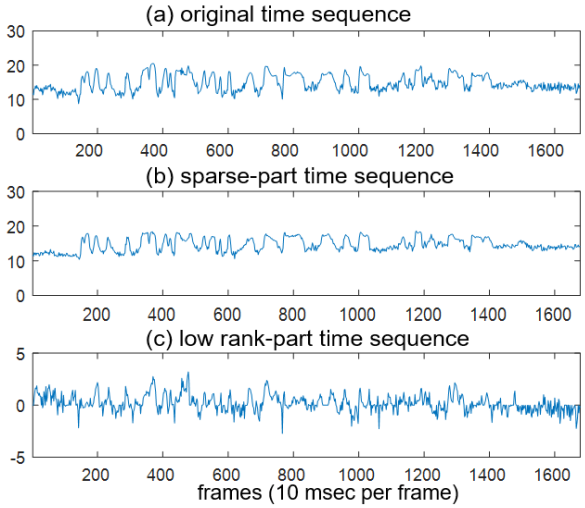


Figure 5. (**a**) The time series of the eighth logarithmic mel-filter-bank coefficients (FBANK) of a noisy utterance; (**b**) The sparse part of (**a**) derived from RPCA; (**c**) The low-rank part of (**a**) derived from RPCA.

4. Experimental Setup

The Aurora-4 database [35] is used to evaluate noise robustness of the features created via the proposed RPCA-SPC. Aurora-4 is a medium vocabulary task (a 5000-word vocabulary task) acquired from the Wall Street Journal (WSJ) corpus [36] at 8 kHz and 16 kHz sampling rates. In Aurora-4, 7138 noise-free clean utterances are recorded with a primary microphone to form the clean-training set, and they are also further contaminated to form the multi-training set with or without the secondary channel distortions and any of six different types of additive noise at the SNRs ranging from 10 to 20 dB. The testing data are split into 14 different test sets (Sets 1–14), with each set containing 330 utterances. The utterances in Sets 1–7 are recorded with a single microphone, while different microphones are used to record the utterances in Sets 8–14, which accordingly contain channel distortions relative to those in Sets 1–7. In addition, Sets 2–7 and Sets 9–14 are further contaminated by additive noise of six types at SNRs from 5 to 15 dB. In particular, as for our experiments we adopt the clean-condition training mode to prepare the acoustic models, and all of the utterances used are at the sampling rate of 8 kHz.

Regarding the speech features, 39-dimensional MFCCs (including 13 static components plus their first- and second-order time derivatives) and 40-dimensional FBANK features serve as the baseline features, and they are further processed by any of mean normalization (MN), mean and variance normalization (MVN), relative spectral (RASTA), and the presented RPCA-SPC. The clean-condition training data is converted to speech features, which are then used to train context-dependent (CD) acoustic models, which further have two different structures, i.e., GMM-HMM and DNN-HMM, in which GMM, DNN, and HMM refer to Gaussian-mixture model, deep-neural network, and hidden Markov model, respectively. Stated in more detail, GMM-HMM and DNN-HMM use GMM and DNN, respectively, to represent each state of the hidden Markov model. As for GMM-HMM, each tri-phone of speech signals and the silence are respectively characterized by a HMM with three states and eight Gaussian mixtures per state and a HMM with three states each having 16 mixtures. On the other hand, seven layers are used for the DNN structure in the DNN-HMMs for tri-phones and silence, having five hidden layers with each layer containing 2048 nodes. A set of trigram language models is created via the reference transcription of training utterances. Finally, the evaluation results are represented using word error rate (WER).

5. Experimental Results and Discussions

5.1. Individual Method

To begin with, Tables 1 and 2 show the WER (%) values for the various features including the baseline ones and those processed by any of MN, MVN, RASTA, and RPCA-SPC, with respect to GMM-HMM and DNN-HMM scenarios. From these two tables, we have the following findings:

1. Both MFCC and FBANK (with GMM-HMM and DNN-HMM as the acoustic models, respectively) give very low WERs for the clean noise-free set, i.e., Set 1. However, FBANK with DNN-HMM outperforms MFCC with GMM-HMM by giving even lower WERs, which is in general attributed to the deep learning scheme in DNN.
2. As for the second sub-group, Sets 2–7, the WERs of both MFCC and FBANK are much higher in comparison with those obtained in Set 1. Thus, we observe that noise deteriorates the performance of speech recognition, and MFCC and FBANK are quite vulnerable to noise. In addition, comparing the WERs obtained from Sets 8–14 with those from Sets 1–7, we see that an extra channel distortion results in further degradation of recognition accuracy. Furthermore, unlike the clean noise-free case, FBANK with DNN-HMM behaves almost equally to MFCC with GMM-HMM, implying that the deep learning structure in acoustic modeling does not necessarily benefit the recognition accuracy under adverse environments.
3. For the clean noise-free case, only MN behaves better than the baseline in promoting both MFCC (with GMM-HMM) and FBANK (with DNN-HMM), while the other methods including MVN, RASTA, and RPCA-SPC worsen the recognition accuracy of the baseline features in particular for the case of GMM-HMM. The probable explanation is as follows: Compared with the other methods, MN simply applies a subtraction operation and does the least change to the original features. By contrast, RASTA and RPCA-SPC always remove some components from speech features in Set 1, which does not have any noise distortion. Accordingly RASTA and PRCA-SPC corrupt clean noise-free speech features by diminishing speech-relevant information or introducing extra distortions.
4. For the mismatched Test Sets (sets 2–7 and 8–14), RPCA-SPC is shown to substantially improve the recognition accuracy relative to the baseline, which undoubtedly shows the effectiveness of the newly proposed RPCA-SPC in enhancing noise robustness of speech features. These results also support our claim that extracting the sparse component in noisy speech features can highlight the clean-speech portion and/or reduce the pure noise portion.
5. For the sets containing noise only (sets 2–7) and the sets containing both channel and noise interferences (sets 8–14), all the methods discussed here provide baseline features with significantly better recognition accuracy. For example, in Sets 2–7, RASTA shows around 23% and 21% in averaged accuracy improvement for the cases of DNN-HMM and GMM-HMM, respectively, and RPCA-SPC gives rise to an averaged accuracy improvement of around 25% and 20% for the cases of DNN-HMM and GMM-HMM, respectively. It is also shown that RPCA-SPC behaves better than MN, MVN, and RASTA in reducing the effect of noise and channel distortions on speech recognition for the case of DNN-HMM.
6. For mismatched noise/channel cases, the recognition accuracy achieved by DNN-HMM is consistently superior to that by GMM-HMM, which coincides with the general idea that deep learning techniques benefit speech recognition. The proposed RPCA-SPC is shown to profit speech features under the DNN-HMM scenario in particular, giving lower word error rates than the other methods, and is thus revealed to further capture the speech-related information beyond what the DNN structure does.

Table 1. The WER (%) for different sets achieved by the GMM-HMM acoustic models with MFCC features as the baseline and the features processed by any of MN, MVN, RASTA, and RPCA-SPC.

	Set 1	Sets 2–7	Sets 8–14
Baseline	4.75	51.58	67.92
MN	4.17	32.58	47.98
MVN	5.38	31.59	47.09
RASTA	5.53	30.04	45.72
RPCA-SPC	7.42	31.70	46.30

Table 2. The WER (%) for different sets achieved by the DNN-HMM acoustic models with FBANK features as the baseline and the features processed by any of MN, MVN, RASTA, and RPCA-SPC.

	Set 1	Sets 2–7	Sets 8–14
Baseline	2.97	52.81	68.81
MN	2.62	28.81	44.60
MVN	2.97	26.54	43.12
RASTA	3.27	29.46	45.77
RPCA-SPC	3.79	27.33	38.82

5.2. The Cascade of RPCA with any of the Other Methods

Next, RPCA-SPC is cascaded with any of MN, MVN, and RASTA in order to see if such a connection is additive to provide even better results than each individual component method. Tables 3 and 4 and Figure 6 show the corresponding results associated with the cases of GMM-HMM and DNN-HMM, respectively. Here the notation "A + B" refers to method A followed by method B, while "B + A" refers to method B followed by method A, where A and B refer the names of methods to be cascaded. Notably, these two types of combinations are different since at least one of the component methods is a non-linear operation. Comparing the results shown in Tables 3 and 4 and Figure 6 with those in Tables 1 and 2, several observations can be made:

1. The series connection of RPCA-SPC with either of MN and RASTA behave better than each single component method for Sets 2–7 (containing noise) and Sets 8–14 (containing noise and channel interference). For example, in the case of DNN-HMM, MN + RPCA-SPC, and RPCA-SPC + MN give 25.67% and 30.53% in WER averaged over the 14 Sets, which are lower than those obtained by MN (34.83%) and RPCA-SPC (31.39%). Similarly, RASTA + RPCA-SPC and RPCA-SPC + RASTA give 27.92% and 25.97% in WER averaged over the 14 Sets, which are lower than those obtained by RASTA (35.75%) and RPCA-SPC (31.39%). These results again claim that the newly proposed RPCA-SPC serves as a promising technique that can extract noise-robust components in speech features either preprocessed/post-processed by MN, RASTA or not. In addition, MN + RPCA-SPC provides a relatively low WER value on average for the DNN-HMM scenario, indicating that it is quite suitable for use and development in the deep learning architecture of speech recognition;
2. Unlike MN and RASTA, cascading MVN with RPCA-SPC does not necessarily gives better recognition results than MVN and RPCA-SPC alone. One possible reason is that MVN makes the features of all different channels behave like a common random variable with zero mean and unity variance, and thus the subsequent RPCA-SPC tends to view these features to be low-rank so as to discard a significant amount of them, which somewhat harms the recognition accuracy. In addition, For RPCA-SPC preprocessed features, MVN is likely to bring an over-normalization effect and diminishes the respective discriminating capability.

Table 3. The WER (%) for different sets achieved by the GMM-HMM acoustic models with MFCC features as the baseline and the features processed by the cascade of RPCA-SPC and any of MN, MVN, and RASTA.

	Set 1	Sets 2–7	Sets 8–14
Baseline	4.75	51.58	67.92
MN + RPCA-SPC	6.43	28.01	41.43
RPCA-SPC + MN	6.52	29.24	42.16
MVN + RPCA-SPC	5.77	27.85	43.99
RPCA-SPC + MVN	6.48	31.68	49.60
RASTA + RPCA-SPC	7.08	29.08	43.99
RPCA-SPC + RASTA	7.58	28.02	41.51

Table 4. The WER (%) for different sets achieved by the DNN-HMM acoustic models with FBANK features as the baseline and the features processed by the cascade of RPCA-SPC and any of MN, MVN, and RASTA.

	Set 1	Sets 2–7	Sets 8–14
Baseline	2.97	52.81	68.81
MN + RPCA-SPC	3.19	20.69	33.15
RPCA-SPC + MN	3.46	27.63	36.89
MVN + RPCA-SPC	3.53	31.40	46.22
RPCA-SPC + MVN	3.89	25.79	36.81
RASTA + RPCA-SPC	4.76	24.17	34.43
RPCA-SPC + RASTA	4.97	22.47	31.97

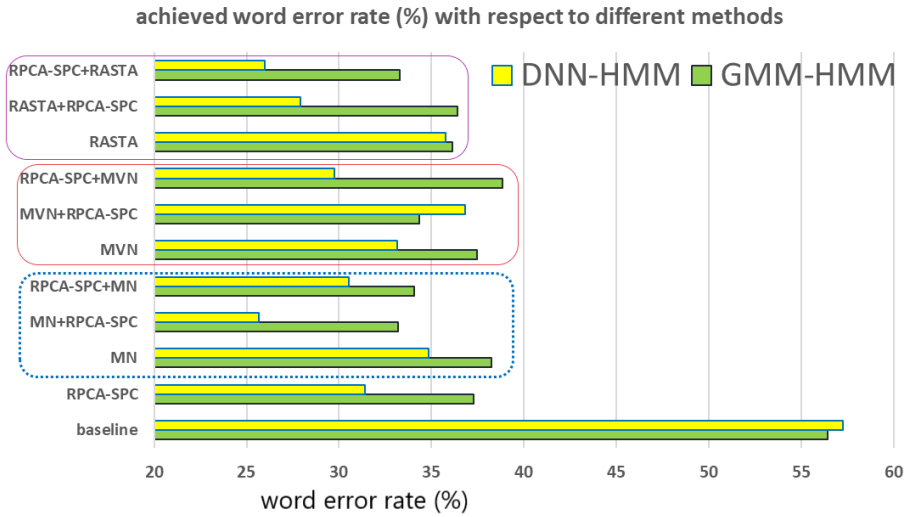


Figure 6. The WER (%) averaged over 14 Test Sets of Aurora-4 for DNN-HMM and GMM-HMM with respect to different robustness methods.

Finally, the WER values averaged over the 14 Test Sets with respect to all the aforementioned methods are depicted in Figure 6. From this figure, it is reconfirmed that the proposed RPCA-SPC benefits the recognition accuracy under adverse environments, and it can be integrated with MN and RASTA, both of which are simple but effective noise-robust methods, to give rise to further better performance. In addition, RPCA-SPC behaves well for both GMM-HMM and DNN-HMM architectures.

5.3. Performing RPCA-SPC on Gammatone-Filterbank Features

In the previous subsections, we have revealed that the proposed RPCA-SPC can enhance noise robustness of FBANK and MFCC features and thus improve the respective recognition accuracy. Here, we would like to further investigate whether RPCA-SPC can be also beneficial to the speech features that are derived from the well-known gammatone filter-bank [3], a widely used model of auditory filters in the auditory system. Briefly speaking, here the creation process of FBANK and MFCC is modified by replacing mel-filters with gammatone filters so as to build the logarithmic gammatone-filterbank coefficients and gammatone cepstral coefficients, denoted by GFBANK and GFCC, respectively. The experiments are conducted almost the same as those described in the previous subsections except replacing FBANK and MFCC with GFBANK and GFCC, respectively. As for the used gammatone filter-bank, the center frequency of the m^{th} filter is determined by the Greenwood function

$$f_c[m] = A\left(10^{am/N} - k\right) \text{ Hz} \tag{4}$$

with $A = 165.4$, $a = 2.1$, $N = 40$ (the number of filters) and $k = 1$, and the frequency response of the m^{th} filter is the magnitude part of the Fourier transform of the impulse response

$$h_m(t) = t^{n-1}e^{-2\pi Bt}\cos(2\pi f_c[m]t + \varnothing) \tag{5}$$

with $n = 4$, $B = 0.025$, and $\varnothing = 0$.

For simplicity, here only the proposed RPCA-SPC is tested, for which results are listed in Figures 7 and 8 for GFCC (with GMM-HMM) and GFBANK (with DNN-HMM), respectively. From these two figures, we reveal that RPCA-SPC significantly improves the recognition accuracy of GFCC and GFBANK features in all mismatched Test Sets (Sets 2–14), while it causes accuracy degradation for the clean noise-free set (Set 1). These results agree well with what we have observed previously in the cases of MFCC and FBANK. Therefore, the effectiveness of RPCA-SPC in enhancing noise robustness of speech features under mismatched conditions is reconfirmed.

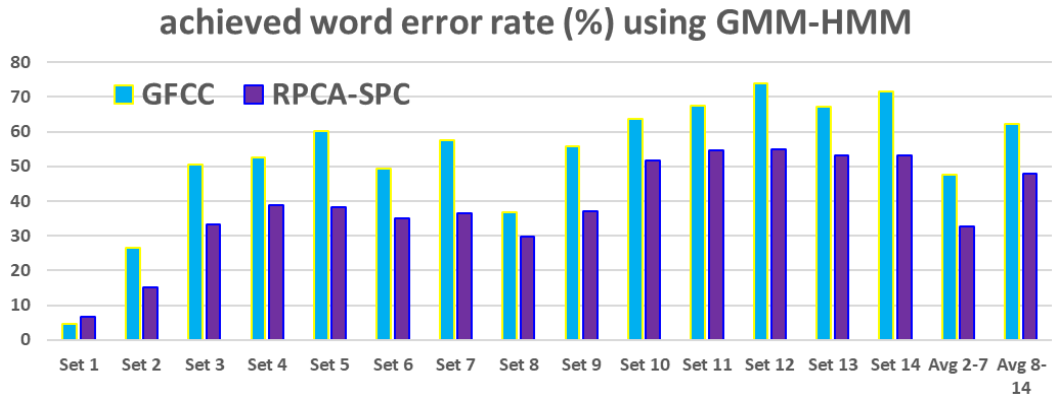


Figure 7. The WER (%) for each of the 14 Test Sets of Aurora-4 for GFCC features using GMM-HMM with respect to the baseline and RPCA-SPC.

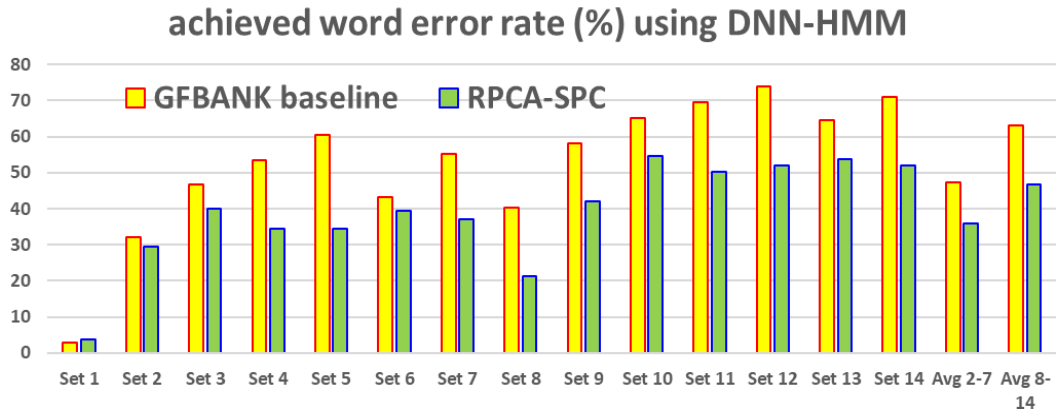


Figure 8. The WER (%) for each of the 14 Test Sets of Aurora-4 for GFBANK features using DNN-HMM with respect to the baseline and RPCA-SPC.

6. Conclusions

In this paper, we propose exploiting robust principal component analysis (RPCA) to extract the sparse partition of the FBANK/MFCC features of noise-corrupted utterances, which corresponds to the speech-dominant elements. The resulting RPCA-wise features show significantly improved recognition performance relative to the original FBANK/MFCC features under the AURORA-4 median-vocabulary recognition task with the renowned deep neural network (DNN) as the acoustic models. Furthermore, the newly proposed features reveal good addition to the popular normalization methods, mean normalization (MN) and relative spectral (RASTA), as the respective integration provides further improvement in recognition accuracy compared with the individual component method. In particular, we do not claim that the proposed RPCA-SPC alone is competitive to state-of-the-art noise-robust algorithms, while it is likely to serve as a pre-processing/post-processing process of other noise-robust methods to achieve better performance. Notably, the acoustic models adopted in the evaluations for RPCA-SPC are either of DNN-HMM and GMM-HMM, while the recent findings reveal that novel deep neural network structures, such as a convolutional neural network (CNN), a recurrent neural network (RNN) and a bi-directional long-short term memory (BLSTM) network, can reduce noise effects significantly for an ASR when they serve as acoustic models. In our opinion, adopting these novel structures alone without any noise-robust feature techniques very likely works well for the ASR in which the training speech data are in a multi-condition manner, i.e., the training set consists of noisy data containing various sources of noise. While an ASR is operated in a clean-condition training scenario, it is less likely that these novel structures can learn the characteristics of noise existing in the application environment. In such a circumstance, a pre-processing stage consisting of noise-robust algorithms, like MN, MVN and the presented RPCA-SPC, for the input speech signals/features is still very beneficial to improve the recognition accuracy of an ASR. In addition, the effectiveness of the presented RPCA-SPC is just evidenced with the Aurora-4 database and task, in which noise-corrupted utterances are created by manually adding noise to clean utterances and thus they are artificial and not real noisy data. However, due to the fact that RPCA-SPC brings about a significant recognition improvement to the Aurora-4 task, it can be reasonably claimed RPCA-SPC is very likely helpful to deal with real noisy data, while this still needs to be further confirmed. As for the future avenue, we will further test RPCA-SPC in real noise-corrupted speech database such as the CHiME-4 dataset [37], investigate whether RPCA-SPC can be well applied to other ASR tasks with different deep learning architectures, and see if it is able to alleviate the reverberant interference in speech signals.

Author Contributions: As for the paper, Prof. J.-w.H. gave the main idea, designed the algorithm and the respective experiments and did the most part of writing. Prof. J.-S.L. served as a consultant and provided precious comments for revising this paper. Mr. P.-J.W. implemented most of the experiments and did some parts of writing.

Funding: This research is partially sponsored by the Ministry of Science and Technology in Taiwan (Grant Number: MOST 106-2221-E-260-006).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.* **1990**, *87*, 1738–1752. [CrossRef] [PubMed]
2. Benesty, J.; Sondhi, M.M.; Huang, Y. *Springer Handbook of Speech Processing*; Springer: Berlin, Germany, 2008.
3. Schlider, R.; Bezrukov, I.; Wagner, H.; Ney, H. Gammatone features and feature combination for large vocabulary Speech recognition. In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, Honolulu, HI, USA, 15–20 April 2007; Volume 4, pp. IV-649–IV-652.
4. Berouti, M.; Schwartz, R.; Makhoul, J. Enhancement of speech corrupted by acoustic noise. In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, Washington, DC, USA, 2–4 April 1979; pp. 208–211.
5. Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal. Process.* **1979**, *27*, 113–120. [CrossRef]
6. Plapous, C.; Marro, C.; Scalart, P. Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Trans. Acoust. Speech Signal. Process.* **2006**, *14*, 2098–2108. [CrossRef]
7. Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal. Process.* **1985**, *33*, 443–445. [CrossRef]
8. Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal. Process.* **1984**, *32*, 1109–1121. [CrossRef]
9. Furui, S. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust. Speech Signal. Process.* **1981**, *29*, 254–272. [CrossRef]
10. Viikki, O.; Laurila, K. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Commun.* **1998**, *25*, 133–147. [CrossRef]
11. Hilger, F.; Ney, H. Quantile based histogram equalization for noise robust large vocabulary speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 845–854. [CrossRef]
12. Lin, S.H.; Chen, B.; Yeh, Y.M. Exploring the use of speech features and their corresponding distribution characteristics for robust speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 84–94. [CrossRef]
13. Hermansky, H.; Morgan, N. RASTA processing of speech. *IEEE Trans. Speech Audio Process.* **1994**, *2*, 578–589. [CrossRef]
14. Xiao, X.; Chng, E.S.; Li, H.Z. Normalization of the speech modulation spectra for robust speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 1662–1674. [CrossRef]
15. Chen, C.P.; Bilmes, J. MVA processing of speech features. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 257–270. [CrossRef]
16. Sun, L.C.; Lee, L.S. Modulation spectrum equalization for improved robust speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 828–843. [CrossRef]
17. Hung, J.W.; Tu, W.H.; Lai, C.C. Improved modulation spectrum enhancement methods for robust speech recognition. *Signal. Process.* **2012**, *92*, 2791–2814. [CrossRef]
18. Hung, J.W.; Hsieh, H.J.; Chen, B. Robust Speech Recognition via Enhancing the Complex-Valued Acoustic Spectrum in Modulation Domain. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 236–251. [CrossRef]
19. Li, J.; Deng, L.; Gong, Y.; Haeb-Umbach, R. An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 745–777. [CrossRef]
20. Leggetter, C.J.; Woodland, P.C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Comput. Speech Lang.* **1995**, *9*, 171–185. [CrossRef]
21. Gales, M.J.F. Maximum likelihood linear transformations for HMM based speech recognition. *Comput. Speech Lang.* **1998**, *12*, 75–98. [CrossRef]
22. Wu, J.; Huo, Q. Supervised adaptation of MCE-trained CDHMMs using minimum classification error linear regression. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Orlando, FL, USA, 13–17 May 2002; Volume I, pp. 605–608.

23. He, X.; Chou, W. Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong, China, 6–10 April 2003; Volume I, pp. 556–559.
24. Yu, K.; Gales, M.J.F.; Woodland, P.C. Unsupervised adaptation with discriminative mapping transforms. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 714–723. [CrossRef]
25. Gales, M.J. Model-Based Techniques for Noise Robust Speech Recognition. Ph.D. Thesis, Cambridge University, Cambridge, UK, 1995.
26. Moreno, P.J. Speech Recognition in Noisy Environments. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 1996.
27. Li, J.; Deng, L.; Haeb-Umbach, R.; Gong, Y.F. *Robust Automatic Speech Recognition: A Bridge. to Practical Applications*; Elsevier: AMS, NL, 2015; Chapter four: Processing in the feature and model domains.
28. Candes, E.J.; Li, X.; Ma, Y.; Wright, J. Robust principal component analysis? *J. ACM* **2011**, *58*, 11. [CrossRef]
29. Bouwmansa, T.; HadiZahzahb, E. Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance. *Comput. Vis. Image Understand.* **2014**, *122*, 22–34. [CrossRef]
30. Sun, C.; Zhang, Q.; Wang, J.; Xie, J. Noise reduction based on robust principal component analysis. *J. Comput. Inf. Syst.* **2014**, *10*, 4403–4410.
31. Wu, C.L.; Hsu, H.P.; Wang, S.S.; Hung, J.W.; Lai, Y.H.; Wang, H.M.; Tsao, Y. Wavelet speech enhancement based on robust principal component analysis. *Proc. Interspeech* **2017**, *781*, 439–443.
32. Gavrilescu, M. Noise Robust Automatic Speech Recognition System by Integrating Robust Principal Component Analysis (RPCA) and Exemplar-Based Sparse Representation. In Proceedings of the International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Bucharest, Romania, 25–27 June 2015.
33. Chen, Z.; Ellis, D.P.W. Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20–23 October 2013.
34. Sun, P.; Qin, J. Low-rank and sparsity analysis applied to speech enhancement via online estimated dictionary. *IEEE Signal Process. Lett.* **2016**, *23*, 1862–1866. [CrossRef]
35. Parihar, N.; Picone, J. *Aurora Working Group: DSR Front. End LVSCR Evaluation au/384/02*; Institute for Signal and Information Processing: Philadelphia, PA, USA, 2002.
36. Paul, D.B.; Baker, J.M. The design for the wall street journal-based CSR corpus. In Proceedings of the workshop on Speech and Natural Language (HLT '91), Harriman, NY, USA, 23–26 February 1992; pp. 357–362.
37. The 4th CHiME Speech Separation and Recognition Challenge. Available online: http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/ (accessed on 14 August 2018).