LEVERAGING JOINTLY SPATIAL, TEMPORAL AND MODULATION ENHANCEMENT IN CREATING NOISE-ROBUST FEATURES FOR SPEECH RECOGNITION

1HSIN-JU HSIEH, 2HAO-TENG FAN, 3JEIH-WEIH HUNG

1,2,3Dept of Electrical Engineering, National Chi Nan University, Taiwan, Republic of China
E-mail: 1s101323902@mail1.ncnu.edu.tw, 2s99323904@ncnu.edu.tw, 3jwhung@ncnu.edu.tw

Abstract- This paper presents to adopt various fusion types of spatial, temporal and modulation domain speech feature enhancement techniques in order to achieve superior speech recognition performance under noise-corrupted environments. With the mel-frequency cepstral coefficients (MFCC) as the standard speech feature representation, the spatial-domain techniques involve the short-time intra-frame feature enhancement, while the temporal-domain techniques compensate for the noise distortion that exists in the long-term inter-frame MFCC time stream. Furthermore, the modulation-domain techniques are conducted on the Fourier transform of a MFCC time stream. The evaluation experiments conducted on the connected-digit Aurora-2 database reveal that each of the spatial/temporal enhancement techniques adopted here performs better than the unprocessed MFCC baseline, and the integration of the methods respectively for spatial-, temporal- and modulation-domain features can result in even better recognition accuracy than the individual component method under a wide range of noise-corrupted environments. These results clearly demonstrate that the methods in the three domains treat noise in different aspects and therefore they are complementary to each other.

Keywords- Noise Robustness, Speech Recognition, Spatial Processing, Temporal Processing, Modulation Domain.

I. INTRODUCTION

Most of the state-of-the-art automatic speech recognition (ASR) systems can perform well in a controlled laboratory environment. However, their performance usually degrades dramatically when they are applied outside the laboratory and in real-world applications. The performance degradation is often caused by interfering sources and distortions which are usually termed the environment variability. This variability and the resulting environmental mismatch between the developing and application situations may be caused by additive noise, channel distortion, different speaker characteristics, etc. In order to alleviate this mismatch, a great number of robustness algorithms have been proposed and thereby the application field of speech recognition can be broadened. These robustness algorithms can be roughly classified into three schools: signal enhancement, feature compensation and model adaptation. First of all, as for signal enhancement, the aim is to improve the quality and intelligibility of speech signals. The corresponding techniques include spectral subtraction (SS) [1]-[3], short-time spectral amplitude estimation based on minimum mean-squared error criteria (MMSE-STSA) [4], MMSE-based log-spectral amplitude estimation (MMSE log-STSA) [5], Wiener filtering [6, 7], Kalman filtering [8], modulation spectral subtraction (ModSpecSub) [9] and minimum mean-square error short-time spectral modulation magnitude estimator (MME) [10], just to name a few.

Next, the general purpose of feature compensation methods is to build a speech feature representation that is robust to noise, and most of these methods are focused on refining the conventional speech features, such as linear predictive coefficients (LPC) [11], mel-frequency cepstral coefficients (MFCC) [12] and perceptual linear prediction (PLP) [13], which behave well in clean noise-free situation, but are vulnerable to noise/interference. One primary direction of this category of methods is to compensate the statistics of temporal feature streams, and several popular feature statistics compensation methods include cepstral mean subtraction (CMS) [14], mean and variance normalization (MVN) [15], cepstral histogram equalization (CHN) [16], higher order cepstral moment normalization (HOCMN) [17] and cepstral shape normalization (CSN) [18].

Finally, the last school of methods, including parallel model combination (PMC) [19], speech and noise decomposition (SND) [20], vector Taylor series (VTS) [21], maximum a posteriori (MAP) [22], maximum likelihood linear regression (MLLR) [23], statistical re-estimation (STAR) and maximum mutual information (MMI) [24, 25], etc., is focused on tuning the acoustic models in the recognizer with respect to noise conditions in application. These methods take into account the noise characteristics within recognition procedures rather than eliminate the noise effect in the input signals/features.

In recent years, our research group has been focused on developing noise-robustness techniques which primarily fall in the category of feature compensation as mentioned earlier. In particular, these developed techniques are to enhance the widely used MFCC speech features on different perspectives, which refer to temporal, spatial and modulation domains. Therefore, in this paper we focus on exploring the effectiveness as for the pairing of any two developed techniques that dwell in different domains and investigating whether such a pairing result in better
performance than each individual component technique. According to the recognition experiments conducted on the well-known Aurora2 database and task [26], it is found that in most cases the noise robustness algorithms in different domains benefit one another and can produce further noise-robust speech features accordingly. These results further show that our presented algorithms in different domains deal with different traits of noise effect, and thereby using them together can further alleviate the degradation on MFCC speech features caused by noise.

The remainder of the paper is organized as follows: Section II briefly reviews the various noise robustness algorithms in three different domains which we have presented. Experimental setup is provided in Section IV, and Section V gives the detailed experimental results for the various integrations of any two algorithms, together with the corresponding discussions. Finally, Section VI contains a concluding remark and future works.

II. REVIEW OF THE ROBUSTNESS ALGORITHMS

2.1 Spatial-domain methods
The spatial-domain methods mentioned here take into consideration the mutual correlation among the intra-frame MFCC features, and the associated algorithms we have developed in [27] are the weighted spatial-MVN and spatial-HEQ, abbreviated by WS-MVN and WS-HEQ, respectively. Briefly speaking, WS-MVN and WS-HEQ adopt the idea of S-HEQ [28] and divide the MFCC features within every individual frame into low- and high-subbands. Then the two sub-bands are weighted according to their relative influence in recognition accuracy, and finally the weighted subbands of MFCC time sequences are enhanced by either MVN or HEQ. We have shown in [27] that WS-MVN and WS-HEQ behave better than S-HEQ and provide promising results in promoting robustness of MFCC features, and they reveal good additive property when applied together with some well-known robustness algorithms like MVN plus ARMA filtering (MVA) [29] and temporal structure normalization (TSN) [30].

2.2 Temporal-domain methods
As pointed out in the introduction section, the well-known CMS, MVN and CHN process the temporal MFCC sequences by compensating the associated statistics. Here we would like to introduce three other temporal processing algorithms that we developed in [31, 32], which are cepstral wavelet denoising (WD), sub-band temporal MVN (SB-TMVN) and sub-band temporal HEQ (SB-THEQ). All of these three algorithms take advantage of the discrete wavelet transform (DWT) to split each cepstral temporal sequence into several sub-bands (approximation and detail parts). Then WD applied a thresholding scheme to remove the relatively small-valued components in each individual sub-band, while SB-TMVN and SB-THEQ adopt a statistics compensation procedure to normalize the mean, variance or histogram of each sub-band temporal sequence. It has been shown that the SB-TMVN and SB-THEQ outperform their full-band counterparts, viz. MVN [15] and CHN [16], and WD behaves better than the conventional wavelet threshold denoising algorithm that operates on the speech waveform directly.

2.3 Modulation-domain methods
Via applying the Fourier transform to the MFCC temporal sequence, the corresponding modulation spectrum can be obtained. The noise effect can be clearly observed in the cepstral modulation spectrum, and thus the respective modulation-domain robustness methods are developed to compensate the modulation spectrum directly. Our recent research have come up with a series of robustness algorithms...
in modulation domain, and some of them are sub-band modulation spectral MVN (SB-MSMVN) [33], sub-band modulation spectral HEQ (SB-MSHEQ) [33] and modulation spectrum power law expansion (MSPLE) [34]. Briefly speaking, SB-MSMVN and SB-MSHEQ split the magnitude component of cepstral modulation spectrum into several segments (i.e., sub-bands) first, and then employ MVN and HEQ to compensate the statistics for each segments. Besides, MSPLE applies a power law operation to the entire magnitude modulation spectrum in order to highlight the lower frequency components that are commonly viewed to be more beneficial for the speech recognition than the higher frequency components. SB-SMVN and SB-SHEQ have been shown to outperform their full-band counterparts, and we have demonstrated that a simple power-law operation as in MSPLE can improve the recognition accuracy significantly.

III. EXPERIMENTAL SETUP

The efficacy of the series of the integrations for different robustness methods mentioned in Section II was evaluated on the noisy Aurora-2 [26] database. Briefly speaking, Aurora-2 is a subset of the TIDIGITS, which consists speech signals uttered by US adults. The task associated with Aurora-2 is to recognize connected digit utterances interfered with various noise sources at different signal-to-noise ratios (SNRs). In the mode of clean-condition training plus multi-condition testing, the acoustic models are trained via 8,440 clean noise-free utterances, and the testing data is further divided into three Sets: Test Sets A and B contain the utterances corrupted by additive noise, and Test Set C is composed of the utterances with additive noise and channel distortion. There are eight noise types in total, and two channel characteristics. Furthermore, the acoustic model for each digit in the Aurora-2 task is set to a left-to-right continuous density HMM with 16 states, each of which is a3-mixture GMM. In regard to speech feature extraction, each utterance of the training and testing sets was represented by a series of 13 static features augmented with their first- and second-order delta coefficients, resulting in a 39-dimensional MFCC feature vector. The training and recognition tests used the HTK recognition toolkit [35], which followed the setup originally defined for the ETSI evaluations [26].

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 The paring of temporal- and modulation-domain methods

At the outset, we evaluate the mode that the original MFCC features are first enhanced by the well-known temporal domain method, temporal MVN (T-MVN), and then further processed by the presented modulation domain methods, SB-MSMVN, SB-MSHEQ and MSPLE. The resulting accuracy rates are summarized in Figure 2. From this figure, several findings can be made:

1. T-MVN brings significant accuracy improvement over the baseline. However, it behaves worse than SB-MSMVN (77.96% in accuracy) and SB-SHE (83.85%).
2. With the T-MVN as the pre-processing method, the performance of SB-MSMVN and MSPLE are further promoted, while SB-SHE drops lightly possibly due to over normalization.
3. Adding MSPLE to T-MVN benefits T-MVN a lot by providing an absolute accuracy improvement of 7.16%. This result implies that the method T-MVN plus MSPLE is well suited in application due to its computational efficiency together with good performance.

![Figure 2. The averaged recognition accuracy rates for one temporal-domain method, T-MVN and three modulation-domain methods, MSPLE, SB-MSMVN and SB-MSHEQ, together with some possible types of integration.](image)

![Figure 3. The averaged recognition accuracy rates for one spatial-domain method, WS-HEQ and three temporal-domain methods, WD, SB-TMVN and SB-THEQ, together with some possible types of integration.](image)

4.2 The paring of spatial- and temporal-domain methods

Next, we present the integration of spatial-domain method, WS-HEQ, and either of three temporal domain methods, WD, SB-TMVN and SB-THEQ. The corresponding evaluation results are shown in Figure 3. From this figure, we have several observations:

1. Any of the various integrations gives rise to an additive effect and show superior accuracy rates in comparison with any individual component.
4.3 The paring of spatial- and modulation-domain methods

Finally, the performance of the fusion of WS-HEQ and either of three modulation domain methods, SB-MSMVN, SB-MSHEQ and MSPLE, is explored. The respective recognition accuracy rates averaged over all noise types and levels in three test sets are shown in Figure 4. This figure reveals that:

1. Most of the combinative procedures produce better results than the individual component method. For example, the integration of WS-HEQ and SB-MSMVN (85.72% in accuracy) behave better than the single WS-HEQ (84.99%) and SB-MSMVN (77.96%).

2. When the feature sequences are pre-processed by WS-HEQ, the three modulation domain methods used here behave very close to each other. This implies that MSPLE is a better choice among the three methods for integrating with WS-HEQ, since it has the simplest computation and achieves nearly optimal recognition accuracy in the integration.

3. Compared with the two sub-band modulation-domain methods, MSPLE only processes the lower band and still can promote the recognition performance as well. In particular, further compensating the WS-HEQ processed features with MSPLE can result in better performance, again evidencing the robustness capability of MSPLE.

CONCLUSIONS AND FUTURE WORKS

In this paper, we have demonstrated the preferable recognition performance of the integration of the robustness methods with respect to different domains for MFCC features. Most types of the integration are simple in implementation and very applicable in real-world scenarios. In the near future, we are going to adopt the speech features enhanced by the presented architectures in the model of state-of-the-art deep-learning neural network (DNN) to evaluate the respective performance. Another direction is to alter the implementation order of the component methods in the integration to see the corresponding effect since these methods are mostly non-linear operations to the original features.

REFERENCES


Leveraging Jointly Spatial, Temporal And Modulation Enhancement In Creating Noise-Robust Features For Speech Recognition

19