IEIE Transactions on Smart Processing and Computing, vol. 3, no. 5, October 2014
http://dx.doi.org/10.5573/IEIESPC.2014.3.5.259

Noise Robust Automatic Speech Recognition Scheme with Histogram of Oriented Gradient Features

Taejin Park, SeungKwan Beack, and Taejin Lee

Audio Research Laboratory, Electronics and Telecommunications Research Institute / Daejeon, South Korea
{inctrl, skbeack, tjlee}@etri.re.kr

***** Corresponding Author: Taejin Park

Received January 15, 2014; Revised March 17, 2014; Accepted July 28, 2014; Published October 31, 2014

** Regular Paper*

***Abstract*:** In this paper, we propose a novel technique for noise robust automatic speech recognition (ASR). The development of ASR techniques has made it possible to recognize isolated words with a near perfect word recognition rate. However, in a highly noisy environment, a distinct mismatch between the trained speech and the test data results in a significantly degraded word recognition rate (WRA). Unlike conventional ASR systems employing Mel-frequency cepstral coefficients (MFCCs) and a hidden Markov model (HMM), this study employ histogram of oriented gradient (HOG) features and a Support Vector Machine (SVM) to ASR tasks to overcome this problem. Our proposed ASR system is less vulnerable to external interference noise, and achieves a higher WRA compared to a conventional ASR system equipped with MFCCs and an HMM. The performance of our proposed ASR system was evaluated using a phonetically balanced word (PBW) set mixed with artificially added noise.

***Keywords*:** Automatic speech recognition, Histogram of oriented gradient, Support vector machine

1. Introduction

Recent developments in automatic speech recognition (ASR) have achieved an almost perfect word recognition rate (WRA), particularly for isolated words. On the other hand, despite their accuracy for recognition tasks in a quiet environment, in a noisy environment, ASR systems perform far below the WRA achieved by the human ear [1]. This suggests ASR systems require more development before they can overcome the effects of noise interference. However, an ASR system with noise robustness is also needed by both the market place and industry. In real-life environments, for example, inside a moving car, voice commanders have difficulty working properly because of the noise interference from the car engine and air friction [2].

The most widely used ASR technique employs Mel-frequency cepstral coefficients (MFCCs) for feature extraction and a hidden Markov model (HMM) for classification [3]. While this approach has been considered the most reliable for ASR tasks, a feature extraction scheme using MFCCs causes a significant mismatch between the features of the clean speech data and the features of the noisy test speech data because of the logarithmic function used in an MFCC. Thus, to reduce this mismatch, a considerable number of studies have been conducted on achieving a feature processing approach.

First, the most widely used technique for increasing the WRA is the MFCC delta-delta (MFCCDD) approach [4]. By adding an additional difference and a double-difference of an MFCC, the MFCCDD feature allows significant improvements in the WRA. A number of feature compensation techniques based on MFCCs have been proposed. Cepstral mean normalization (CMN) [5] is a widely used technique for reducing the mismatches between the test data and training data by subtracting the offset of the mean of the cepstral values. Another popular method, mean variance normalization (MVN), is a widely used approach for increasing the WRA [6]. Moreover, a histogram equalization (HEQ) technique was proposed to minimize the gap between the training and test data [7].

Several attempts have been made to employ other feature methods. Relative spectral perceptual linear prediction (RASTA-PLP) has been employed to achieve noise robust ASR by employing the concept of psychophysics in feature extraction [8, 9]. In addition,

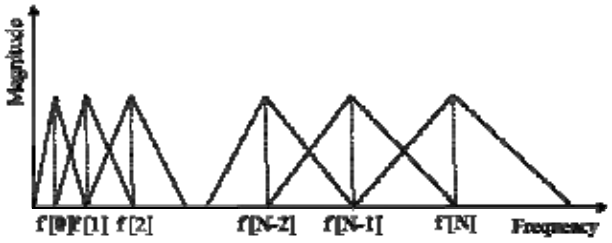


Fig. 1. Mel-scale filters.

there was a recent attempt to employ physiologically inspired spectro-temporal Gabor filter banks (GBFB) to the feature extraction process and improve the MFCCDD baseline [10].

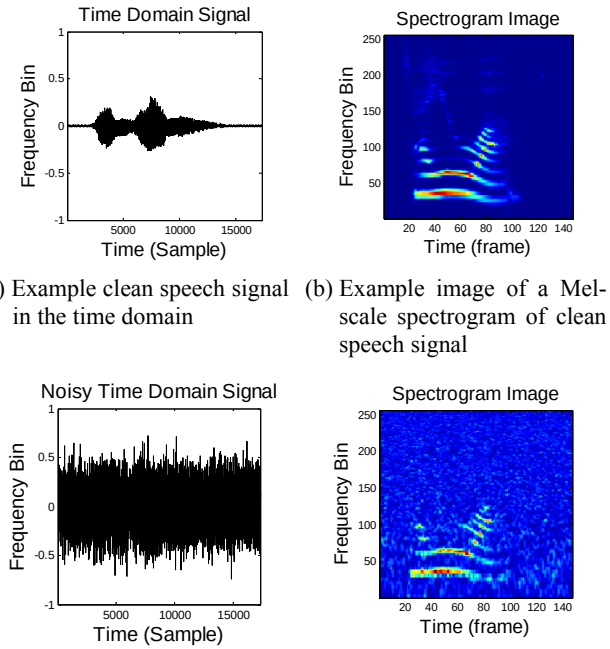
The proposed ASR system, however, uses a different approach by employing a histogram of oriented gradient (HOG) method, which was inherited from image detection technology [11]. The HOG approach showed the best performance with regard to human image detection, demonstrating an almost perfect recognition rate. In contrast, we employed this technique to ASR tasks by considering a spectrogram image as an image to recognize.

Similar to the human image recognition task in [11], the ASR tasks with noise interference can be considered as an image recognition task of desired object image and background. To illustrate this, Fig. 2 shows that, if a spectrogram image of a noise signal is considered as a background image, a speech signal can also be considered as a desired object image for an ASR task.

The strength of the proposed approach is in the way HOG feature extracts gradients from spectrogram image. Since HOG feature considers both horizontal and vertical gradients, HOG feature can extract the both temporal and spectral information from spectrogram images. Since speech signal is clearly visible in spectrogram images, as Fig. 2 describes, even with severe noise level, we thought that HOG feature can extract feature from speech signal even in noisy environment.

However, ASR is a task for classifying the target image, whereas the original purpose of the HOG technique in [11] is detection. Therefore, there is a radical difference between an ASR task and the human detection task described in [11]. Moreover, we also applied a different classifier based on a Support Vector Machine (SVM), which is rarely used in ASR tasks. As mentioned above, most ASR systems use an HMM as a classification scheme. Nevertheless, our proposed feature is unable to work along with an HMM classifier because of the characteristics of the proposed image gradient based feature. Thus, we focused on a different classification scheme, and found that an SVM achieves the best performance. However, the implementation of an SVM is not a concern of the present research. The purpose of this paper is designing an effective feature extraction system that is robust to external noise. In addition, in this paper, an MFCC-based feature was also tested using an SVM classifier to verify the performance of MFCC-based feature extraction scheme as compared to our proposed feature extraction scheme.

The proposed ASR system was evaluated using a phonetically balanced word (PBW) set of Korean words.



(a) Example clean speech signal in the time domain

(b) Example image of a Mel-scale spectrogram of clean speech signal

(c) Example noisy speech signal in the time domain

(d) Example image of a Mel-scale spectrogram of noisy speech signal

Fig. 2. Example of a clean speech signal and the Mel-scale spectrogram for the clean signal.

Using an identical test corpus, the test speech signals contained artificially mixed real-life noise and additive white Gaussian noise. The recognition performance of the proposed ASR system shows that it can recognize words in a noisy environment more accurately.

The remainder of this paper is structured as follows. Section 2 describes the image formation process before moving on to the feature extraction process. Section 3 describes the image mapping process for enhancing the recognition performance. Section 4 details the proposed feature extraction process. The classification of the feature set is mentioned briefly in Section 5. Section 6 reports the experimental results. Section 7 provides the conclusions.

2. Spectrogram Image Generation

For the proposed scheme, which is based on an image detection technique, it is crucial to transform an input speech signal into a proper spectrogram image that can be recognized and classified adequately using the classifier in our proposed ASR system. Similar to a conventional speech processing technique, we applied a Mel-scale into the spectrogram generating process. After a discrete Fourier transform (DFT) process, the scaling process is applied. Fig. 1 describes the frequency domain scaling approach using a Mel-scale [12] scaled to the human auditory system. Each center frequency in Fig. 1 follows the equation below

$$f[k] = 700\left(10^{m[k]/2595} - 1\right) \quad (1)$$

where k is index of filter and $m[k]$ is a Mel-scale number.

By employing a Mel-scale in a spectrogram, we can extract more information from a lower frequency using a limited number of pixels. In addition, to reduce the dimension of the final pattern vector, the time axis is linearly scaled. Therefore, the final image is scaled to a 256 pixel x 128 pixel image. An example spectrogram image of a speech signal is shown in Figs. 2(b) and (d).

3. Mapping process

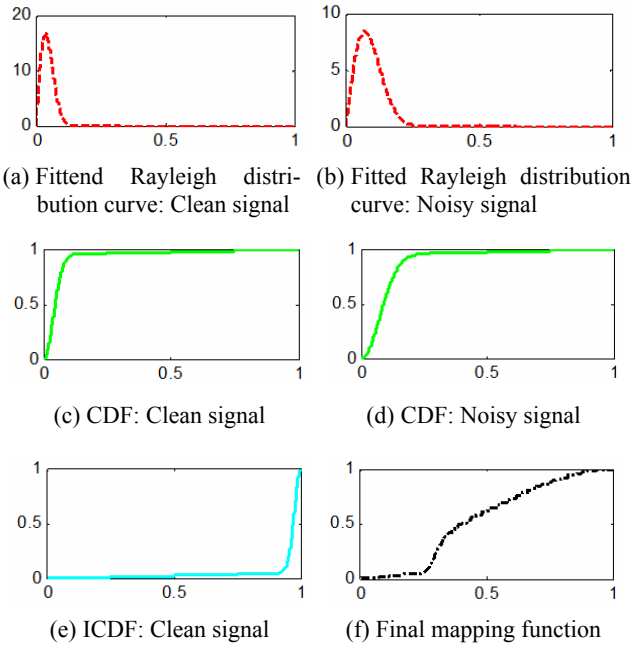
Because our system employs an image recognition technique, a spectrogram image should contain proper information to recognize and classify the speech signal exactly. If a speech signal with environmental noise is recorded, it will directly appear in the spectrogram image. Therefore, like other conventional ASR systems, interference noise deteriorates the WRA of our proposed system. Although there are various noise reduction schemes for increasing the WRA, we propose another method for improving the WRA using an image-processing related technique.

In the field of image processing, there have been innumerable works have attempted to enhance the image quality. One way to enhance the image quality is by modifying the histogram of the pixel intensity values of the image. This technique was originally applied to photographs taken under insufficient light. Photographs with adequate light usually form an equally distributed shape of the pixel intensity value histogram. Therefore, by mapping each pixel of the image into an equally distributed curve, we can obtain a visually well-balanced and clear image.

A similar idea can also be applied to spectrogram images of a speech signal. A histogram curve from the spectrogram image of a speech signal, without the presence of interference noise, usually forms a curve similar to a Gamma distribution [13]. Because determining this distribution is the key to successfully classifying and processing speech signals, many studies have attempted to find the parameters of distribution that are similar to the histogram from the actual spectrogram of a speech signal [14]. Although a range of distribution parameters have been proposed to describe the distribution of pixel intensity values in the spectrogram image of a speech signal, for simplicity, we employed parameters that turn a Gamma distribution into a Rayleigh distribution [14]. A Rayleigh distribution can be described through the following equation:

$$g(x;\sigma) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} \quad x \geq 0, \tag{2}$$

where x is the magnitude of each pixel from a spectrogram image, and σ is the scaling factor of the Rayleigh distribution. The spectrogram images of clean speech signals generally have in common pixel intensities with a Rayleigh distribution and a fairly regular value of scale factor σ_0. On the other hand, if an input signal for the ASR



(a) Fittend Rayleigh distribution curve: Clean signal (b) Fitted Rayleigh distribution curve: Noisy signal

(c) CDF: Clean signal (d) CDF: Noisy signal

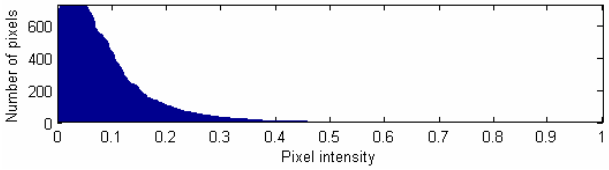
(e) ICDF: Clean signal (f) Final mapping function

Fig. 3. Curve of the Rayleigh distribution fitted to the distribution of pixel intensity values in a spectrogram image of (a) a clean signal and (b) a noisy signal. Figure (c) is CDF of (a), and Figure (d) is CDF of (b). Figure (e) is Inverse CDF of (a). Figure (f) denotes the final mapping curve.

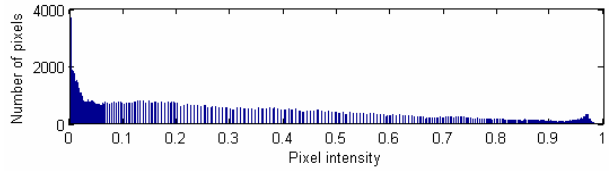
system experiences interference from additive noise, this distribution no longer fits the Rayleigh distribution with a scale factor of σ_0 because the scale factor, σ_0, of a Rayleigh distribution increases from the additive noise level existing in the pixel intensity. This is described in Fig. 3. In noisy speech (Fig. 3(b)) a fitted Rayleigh distribution curve has a higher scale factor compared to that of clean speech (Fig. 3(a)). A scale factor with a higher σ_0 usually indicates that the spectrogram image contains some aspects irrelevant to a speech signal. Therefore, we should map this value back to the original scale factor σ_0 to suppress the effect of the noise artifacts.

However, as we mentioned above, in an image processing technique, the distribution of pixel intensities is equalized to obtain a balanced image contrast. On the other hand, in our proposed technique, we map the histogram-equalized spectrogram image was mapped again to make the scale factor of the final distribution with σ_0, thus leading to a reduction in the mismatch between the original speech signal because the training data usually contain a clean speech signal. We therefore need to equalize the histogram of the spectrogram image first, and then map the resulting image again to an image with scale factor σ_0.

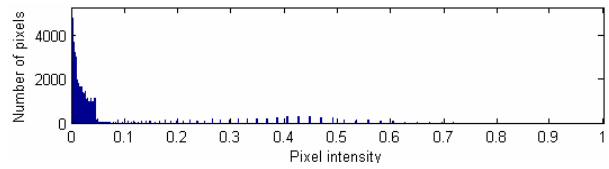
A well-known strategy for equalizing the histogram of pixel intensity values is a method employing a cumulative distribution function (CDF) of the original distribution as a equalization function [15]. To obtain the CDF, a curve fitted of the histogram curve of the pixel intensities of the input image should be obtained in advance by fitting the



(a) An original histogram



(b) An equalized histogram



(c) A final mapped histogram of noisy speech

Fig. 4. Histogram for each process phase.

histogram into a Rayleigh distribution using the scale factor σ_n. Then, equalization function, $E(x)$, which is the CDF of Eq. (2), for the histogram of the pixel intensity values of a noisy speech signal can be derived as follows:

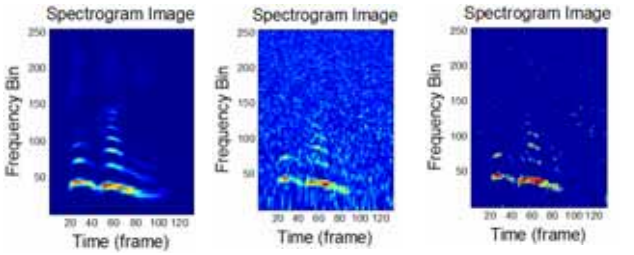
$$E(x;\sigma_n) = 1 - e^{-\frac{x^2}{2\sigma_n^2}} \quad x \geq 0, \tag{3}$$

where x is the intensity value of each pixel from the spectrogram image. An example of Eq. (3) is also shown in Fig. 3(d). After equalizing the histogram, the histogram-equalized spectrogram image should be mapped again through the inverse CDF (ICDF) of the Rayleigh distribution using the regular scale factor σ_0 of a clean signal in order to make the pixel distribution similar to the distribution of the pixel intensity values of the clean signal.

First, using the CDF of the pixel intensity value distribution of a clean speech signal with the original scale factor, σ_0, $M(x)$ can be described through the following equation.

$$M(x;\sigma_0) = 1 - e^{-\frac{x^2}{2\sigma_0^2}} \quad x \geq 0 \tag{4}$$

Second, an example of an inversed version of Eq. (4), which is an ICDF, is described in Fig. 3(e). Accordingly, using an inversed mapping function, $M^{-1}(x)$, an equalized spectrogram image can be mapped back to a spectrogram image containing pixel intensity values following a Rayleigh distribution with a scale factor σ_0. The whole mapping function can be defined as



(a) Spectrogram image of clean speech signal

(b) Spectrogram image of noisy speech signal before mapping

(c) Spectrogram image of noisy speech signal after mapping

Fig. 5. A difference of spectrogram image: before and after the mapping process.

$$C(x) = M^{-1}(E(x,\sigma_n),\sigma_0), \tag{5}$$

where $C(x)$ is the final mapping function. An example of Eq. (5) is shown in Fig. 3(f).

Examples of the histogram equalization and re-mapping process are shown in Fig. 4 along with the histogram graph itself. Fig. 4(a) shows the original histogram of the pixel intensity values of a noisy input speech signal. After the equalization process, the histogram of the input signal is equally distributed throughout a zero-to-one pixel intensity scale, as shown in Fig. 4(b). Finally, after the re-mapping process using the ICDF of Eq. (4), the histogram is rearranged as shown in Fig. 4(c).

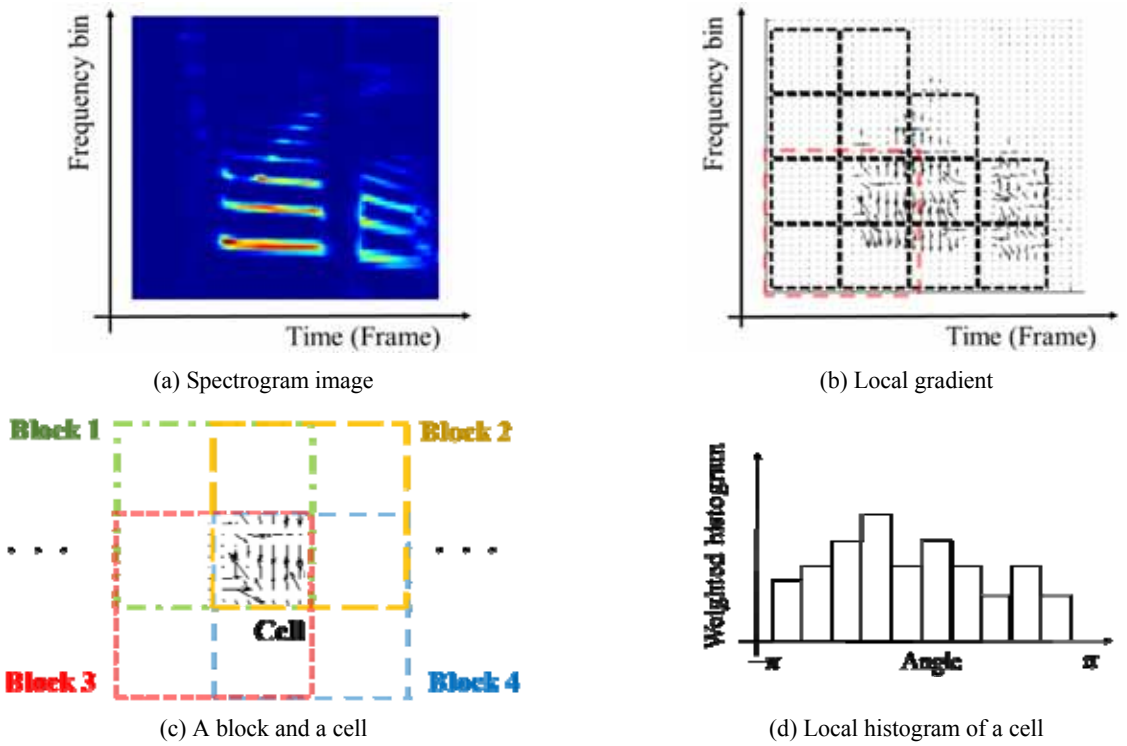
The effects of the histogram equalization and re-mapping are also distinctly shown in the output spectrogram image. As described in Fig. 5, the mapped spectrogram image shows weak repressed pixels and enhanced salient pixels. Although the output image does not perfectly match the original spectrogram image, which is described in Figs. 5(a) and (c), this mapping approach enhances the WRA of our proposed ASR system. Details regarding the effect of mapping on the proposed ASR system are provided in chapter 6.

4. Feature Extraction Scheme

Unlike a conventional feature extraction technique, our proposed ASR system employs the gradient of a spectrogram image. Our approach for gradient extraction generally follows the method proposed in [11]. To achieve this, the gradient of each pixel should be obtained in advance of the feature extraction process. The gradient of each pixel is calculated using the 2-D convolution of the image with the following mask matrix.

$$\mathbf{g} = [-1,\ 0,\ 1] \tag{6}$$

As a result, this mask matrix calculates the difference between the pixels around the center of the 3 x 3 matrix both horizontally and vertically. Therefore, we obtain



(a) Spectrogram image (b) Local gradient

(c) A block and a cell (d) Local histogram of a cell

Fig. 6. (a) A spectrogram image, (b) local gradients, (c) a concept of blocks and cells, (d) an example of a local histogram for a cell.

matrix dT, which contains differences in the time domain, and matrix dF, which contains differences in the frequency domain as follows:

$$\begin{aligned} \mathrm{dT} &= \mathbf{g} \otimes M \\ \mathrm{dF} &= \text{-}\mathbf{g}^T \otimes M \end{aligned}, \tag{7}$$

where M is the spectrogram image of a speech signal, and $\otimes$ is a 2-D convolution operation. Using these difference data, we obtain the angle and magnitude information of the spectrogram image through the equations described below:

$$\begin{aligned} \theta(t,f) &= \arctan\left(\frac{\mathrm{dF}(t,f)}{\mathrm{dT}(t,f)}\right) \\ A(t,f) &= \sqrt{\mathrm{dF}(t,f)^2 + \mathrm{dT}(t,f)^2} \end{aligned}, \tag{8}$$

where $\theta(t,f)$ is a matrix containing the angle information, and $A(t,f)$ is a matrix containing the magnitude information, which form a local gradient. If we set the number of angle bins as N, an N x 1 local histogram column matrix h is calculated from a cell through the following description:

$$h(i) = \sum_{\theta(t,f)\in B(i)} A(t,f), \tag{9}$$

where $h(i)$ is a column vector containing a weighted histogram, which is described in Fig. 6(d), and $B(i)$ denotes the angle bins divided equally from $-\pi$ to π. After calculating these local gradients, they are grouped into a box called a cell. The concept of a cell is shown in Figs. 6(a)-(c). A cell contains an 8 x 8 matrix containing 64 gradient pixels, and a block comprises four cells, as shown in Fig. 6. Four of the local histograms of each cell form a block histogram matrix, as described below:

$$H_B = [h_1^T \; h_2^T \; h_3^T \; h_4^T]^T \tag{10}$$

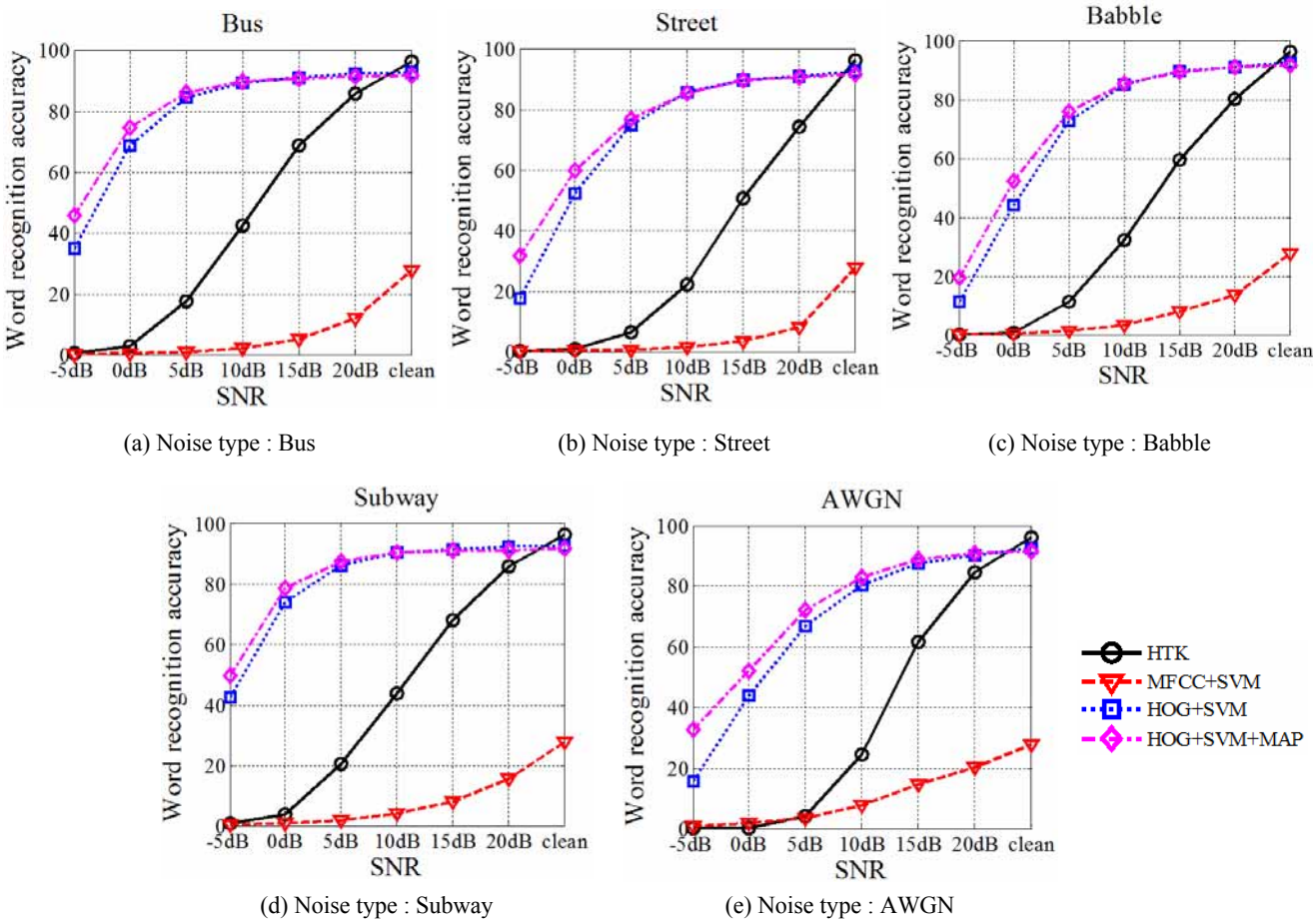
where H_B is a congregated histogram consisting of four column vectors containing local weighted histograms. A block histogram matrix H_B is calculated for every four cells in the image with two cells overlapping each time.

5. Feature Classification

While conventional ASR systems use an HMM as a classifier, we employ a SVM classifier instead, which works well with our given feature extraction technique. However, the classification algorithm used is out of the scope of the present paper, and we employed a technique and implementation that were previously proposed in [16].

The feature vector from Eq. (10) was aggregated again as follows:

$$H_S = [H_{B,1}^T \; H_{B,2}^T \; H_{B,3}^T \; \cdots \; H_{B,M}^T]^T, \tag{11}$$



(a) Noise type : Bus (b) Noise type : Street (c) Noise type : Babble

(d) Noise type : Subway (e) Noise type : AWGN

Fig. 7. Word recognition accuracies for various types of noise.

where M is the total number of blocks in a given input signal. The feature vector H_S is calculated for every speech signal. The SVM classifier learns and classifies the speech signals using these vectors.

6. Experimental Results

To verify its performance, we compared our proposed ASR system with a widely known conventional ASR system called a Hidden-Markov Toolkit (HTK). The settings used for the HTK followed the most widely used settings. To be specific, for the MFCCs and MFCCDD coefficient, 13 cepstral coefficients from a 23 Mel-filter bank were congregated into a single feature vector. Using this MFCC vector set, the delta and double-delta of the MFCCs were attached to the MFCC vector set. Thus, a total of 39 dimensional features were used for the HTK setting. With these features, a mono-phone based HMM model with a 5-mixture model was employed for the recognition test. In addition, a cepstral mean subtraction [5] was applied to the HTK system to maximize the performance of the WRA of the HTK-based ASR system. For the window size, a window length of 480 samples with a window shift of 160 samples was used.

On the other hand, for our proposed ASR system, a 512-sample window with a hop size of 128 samples was used. The spectrogram image size was 128 x 48, which contains 128 Mel-scaled frequency bins and 48 time frames. For the regular scale factor of the Rayleigh distribution of a spectrogram image from the clean speech signals, we used a regular scale factor σ_0 of 0.025. In addition, the MFCCDD feature with the same SVM classifier combination was also tested to verify the performance of the feature itself. For the test set of the MFCCDD and SVM combination, the same MFCCDD feature specification was used.

All ASR methodologies employed the exact same PBW dataset [17]. This PBW dataset consists of 452 words that were equally selected in terms of phonation. In addition, both males and females with a wide variety of ages are engaged in the data gathering process. However, for our speech recognition task, to level the playing field, both the proposed and HTK baseline methodologies used the same numbers of datasets for training.

To evaluate the performances of the ASR systems in an environment of artificial noises, we employed four kinds of natural noise signals (subway, bus, babble, and city streets). These noise signals were recorded in locations around urban areas [18]. In addition, we tested both the HTK-based ASR system and our proposed ASR system in additive white Gaussian noise.

Fig. 7 shows the overall results of the proposed ASR system (HOG+SVM), the HTK-based system (HTK), and

the combined MFCCDD/SVM system (MFCC+SVM). Compared to the HTK-based ASR system, our proposed system showed meaningful improvement. On average, the proposed ASR system outweighed the HTK-based ASR system by 32.99%. However, the proposed ASR system showed a 3.76% lower performance compared to the HTK-based ASR system under clean conditions, largely because the HTK system analyzes a single word into a short time span, and estimates the most probable word base in connection with a short segment speech signal. This process provides more flexibility to the HTK-based ASR system and leads to a higher recognition accuracy under clean conditions. On the other hand, as shown in Fig. 6, our proposed method extracts features from an image segment that overlaps a large area from other block segments of the image. Thus, the feature unit for our proposed recognition system is larger than that for the HTK-based system, thereby leading to a lower accuracy under clean conditions. However, for a noisy environment, the gradient feature from a large area tends to have more flexibility with additive noise because the direction of the gradient hardly changes even when the signal is somewhat contaminated with external noise. Moreover, unlike with an MFCC, the proposed feature extraction technique extracts information from both the time axis and frequency axis, and thus obtains more stable features from the signal. Accordingly, these characteristics give rise to a higher WRA.

In addition, the mapping technique (HOG+SVM+MAP in Fig. 7) was also shown to bring about a higher WRA, which is on average 2.92% higher than in the ASR system without a mapping procedure because a mapping procedure effectively reduces the gap between the training and test signals. However, the ASR system with a mapping procedure shows a 0.9% lower WRA under clean conditions. The degradation of the WRA caused by the mapping function under a clean environment mainly occurred because the scale factor σ_0 is not perfectly identical for every clean speech signal. This small variance in the scale factors caused a mismatch between the test and training signals, thus leading to slight degradation of the WRA under clean conditions.

Finally, as shown in Fig. 7, the combined system of MFCC-based features and the SVM classifier failed to achieve a reasonable WRA. As mentioned earlier, this largely arose from the characteristics of the SVM classifier. Because the MFCC feature extraction process in the HTK system cuts a speech signal into short segments and matches it with a mixture model, the SVM classifier may be inflexible in classifying the signals. Therefore, MFCC-based features and the SVM classifier are unable to cope with the warping or other internal variances of the speech signals.

To summarize the experimental results, the proposed ASR system showed a significantly improved WRA in a noisy environment compared to the widely used HMM- and MFCC-based ASR systems. Moreover, the proposed mapping technique also improves the WRA in a noisy environment. In contrast, the MFCC-based features were unable to achieve a similar WRA using an SVM classifier.

7. Conclusions

In this paper, an ASR system employing a gradient of angle histogram for the classification features was described. Unlike a conventional approach, we employed a pattern classification technique based on an image recognition technique. By extracting the gradient feature from the spectrogram image of a speech signal, we achieved an improved WRA compared to the baseline system based on an HMM and MFCCs. In addition, a spectrogram image processing technique that maps the histogram of the pixel intensity value of a speech signal was proposed. Through this mapping technique, we further improved the WRA of the proposed ASR system in a noisy environment. However, the proposed method shows a slightly lower WRA in a noise-free environment. The degradation in a clean environment should be addressed in future studies.

As mentioned earlier, speech recognition systems often suffer from noise artifacts that degrade the recognition rate. However, in this paper, we proposed a solution to cope with external noise in an ASR system without the use of a noise cancellation technique. We expect our proposed technique to bring about significant improvements in ASR systems if applied to commercial devices.

References

[1] R. P. Lippmann, "Speech recognition by machines and humans," *Speech communication*, Vol. 22, No. 1, pp. 1-15. 1997. Article (CrossRef Link)

[2] A. Torre, D. Fohr, and J. P. Haton, "On the Comparison of Front-Ends for Robust Speech Recognition in Car Environments," *in Proc. ISCA ITRW on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, 2001, pp. 105-108. Article (CrossRef Link)

[3] G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*. Cambridge: Entropic Cambridge Research Laboratory, 1997. Article (CrossRef Link)

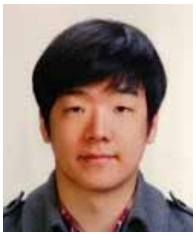
[4] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, Vol. 28, No. 4, pp. 357-366, Aug. 1980. Article (CrossRef Link)

[5] A.E. Rosenberg, C.H. Lee, F. K. Soong, 1994. "Cepstral channel normalization techniques for HMM-based speaker verification," *in Proc. ICSLP*, Vol. 4, pp. 1835-1838, 1994. Article (CrossRef Link)

[6] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, Vol. 25, No. 1-3, pp. 133-147, 1998. Article (CrossRef Link)

[7] A. Torre, et al., "Histogram equalization of speech representation for robust speech recognition," *Speech and Audio Processing, IEEE Transactions on*, Vol. 13, No. 3, pp. 355-366, May. 2005. Article (CrossRef

Link)
[8] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Acoustical Society of America Journal*, Vol. 87, pp.1738-1752, Apr. 1990. Article (CrossRef Link)
[9] H. Hermansky and N. Morgan, "RASTA processing of speech," Speech and Audio Processing, *IEEE Transactions on*, Vol. 2, No. 4, pp. 578-589, Oct. 1994. Article (CrossRef Link)
[10] M. R. Schädler, R. Marc, B. T. Meyer, and B. Kollmeier. "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition." *The Journal of the Acoustical Society of America*, Vol. 131, No. 5, pp. 4134-4151, 2012. Article (CrossRef Link)
[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *in Proc. CVPR*, San Diego, CA, USA, Jun, 2005, pp. 886-893. Article (CrossRef Link)
[12] D. O'Shaughnessy, *Speech communication: human and machine*, Addison-Wesley, 1987, p. 150. Article (CrossRef Link)
[13] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors." *Speech and Audio Processing, IEEE Transactions on*, Vol. 13, No. 5, pp. 845-856, 2005. Article (CrossRef Link)
[14] T. Gerkmann, and R. Martin. "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances." *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, 2010. Article (CrossRef Link)
[15] N. Bassiou, and C. Kotropoulos, "Color image histogram equalization by absolute discounting back-off." *Computer Vision and Image Understanding*, Vol. 107, No. 1, pp. 108-122, 2007. Article (CrossRef Link)
[16] C. C. Chang, and C. J. Lin. "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 27, No. 2.3, 2011. Article (CrossRef Link)
[17] Y.-J Lee, B.-W. Kim, J.-J Kim, O.-Y. Yang, and S.-Y. Lim, "Some considerations for construction of PBW set," *in Proc. of the 12th Workshop on Speech Communications and Signal Processing. Acoustical Society of Korea*, pp. 310-314, Jun. 1995. Article (CrossRef Link)
[18] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on* , vol., no., pp.1,4, 20-23 Oct. 2013. Article (CrossRef Link)



Taejin Park received his B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, Korea, in 2010 and 2012. He has been in Electronics and Telecommunications Research Institute (ETRI) since 2012 as a researcher. His research interests include audio signal processing, speech recognition and machine learning.

Seungkwon Beack received the BS degree in electrical engineering from Hankuk Aviation University, Korea, in 1999, MS and PhD degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2001 and 2005. He has been a senior member of research staff with ETRI, Daejeon, Korea, since 2005. His research interests are in the fields of audio and speech signal processing, spatial audio processing, and multi-channel signal processing.

Taejin Lee received the BS and MS degrees in electronics engineering from Chonbuk National University, Jeonju, Korea, in 1996 and 1998, and his PhD degree in electrical engineering from Chungnam National University, Daejeon, Korea in 2013. He worked for Mobens Co., Ltd. Korea, from 1998 to 2000. He has been in Electronics and Telecommunications Research Institute (ETRI) since 2000, and he is now a Principal Member of Engineering Staff and the Leader of Audio Research Laboratory. From 2002 to 2003, he was a Visiting Researcher at Tokyo Denki University, Japan. His research interests include audio signal processing and interactive broadcasting technologies.