14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

HOUGH TRANSFORM BASED MASKING IN FEATURE EXTRACTION FOR NOISY SPEECH RECOGNITION

Eric H. C. Choi

ATP Research Laboratory
National ICT Australia
Locked Bag 9013, Alexandria, Sydney, NSW 1435, Australia
Eric.Choi@nicta.com.au
www.nict.com.au

ABSTRACT

Despite various advances in recent years, robustness in the presence of various types and levels of environmental noise remains a critical issue for automatic speech recognition systems. This paper describes a novel and noise robust front-end that incorporates the use of Hough transform for simultaneous frequency and temporal masking, together with cumulative distribution mapping of cepstral coefficients, for noisy speech recognition. Recognition experiments on the Aurora II connected digits database have revealed that the proposed front-end achieves an average digit recognition accuracy of 83.31% for all the three Aurora test sets. Compared with the recognition results obtained by using the ETSI standard Mel-cepstral front-end, this accuracy represents a relative error rate reduction of around 57%.

1. INTRODUCTION

State-of-the-art automatic speech recognition (ASR) systems offer good performance if the training and usage conditions are similar and reasonably controlled. However, under the influence of noise, these systems begin to degrade and their accuracies may become unacceptably low in some severe environments. To remedy this noise robustness issue in ASR due to the mismatch in training and usage conditions, various adaptive techniques have been proposed. A common theme of these techniques is the utilization of some form of compensation to account for the effects of noise on the speech characteristics. Typical approaches to improving ASR robustness [1] include pre-enhancing the noisy speech signal [2], feature-space compensation of mismatch between clean and noisy speech features [3], and model-space methods that account for the effects of noise in the speech models [4].

In this work, the main focus is on feature-space compensation for a cepstral based front-end. It is demonstrated that a novel simultaneous frequency and temporal masking derived from Hough transform of the time-frequency spectral sequences of a speech signal can be used together with cumulative distribution mapping to better compensate the effects of additive noises.

The organization of this paper is as follows. It will describe the details of the proposed front-end in Section 2. Following this in Section 3 will be some recognition experiments on the Aurora II digits database and the discussion of the related findings. Finally, a conclusion will be presented in Section 4.

2. PROPOSED FRONT-END

The proposed front-end, with its block diagram shown in Figure 1, is made more robust by incorporating two additional processing modules into the ETSI standard Mel-frequency cepstral coefficient (MFCC) front-end [5]. These new processing modules include Hough transform based masking (HTM) and cumulative distribution mapping (CDM) for the resultant cepstral coefficients.

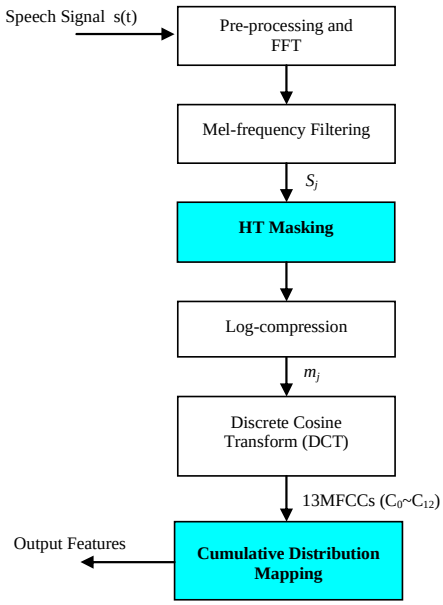

Figure 1- Block diagram of the proposed front-end

Typically, the MFCCs (C_i) of a frame of speech data without any noise compensation are given by:

$$C_i = \sum_{j=1}^{M} m_j Cos[\frac{i\pi}{M}(j-0.5)]; \quad m_j = \log_e(S_j); \quad i = 0,1,2,.....,N; N < M \tag{1}$$

where S_j is the output magnitude of the j-th Mel-filterbank and M is the total number of Mel-filters in the filterbank analysis.

2.1 Spectral Masking using Hough Transform

The Hough transform (HT) is a classical image processing algorithm for extracting parametric patterns, such as lines and circles, from a noisy image [6]. The HT method to extract a straight line in an image is based on transforming the position of a pixel at (x,y) on a line in the image plane (X-Y coordinates) into a sinusoidal curve on the Θ-R plane via:

$$r = x\cos\theta + y\sin\theta \tag{2}$$

where r is the length of a normal to the line from the origin and θ is the angle between this normal and the X-axis. An illustration of the Hough transform is shown in Figure 2. In the figure, the position of an image pixel at (x_1,y_1) is transformed into a sinusoidal curve on the Θ-R plane (the curve with smaller amplitude) and the line on X-Y plane joining (x_1,y_1) and (x_2,y_2) is transformed into a point (θ_o,r_o) on the Θ-R plane.

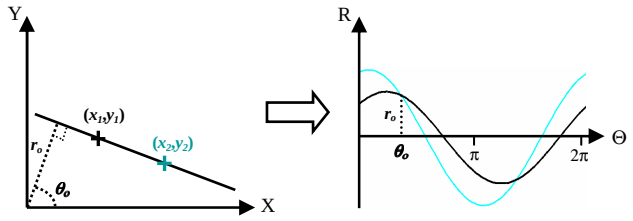

Figure 2 – An illustration of the Hough transform

For an image with intensity value $I(x,y)$ at point (x,y), the most significant straight line in polar-form $(\hat{r},\hat{\theta})$ can be extracted as:

$$(\hat{r},\hat{\theta}) = \arg\max_{r,\theta}\{Acc(r,\theta)\};$$
$$Acc(r,\theta) = \sum_{(x,y)\in L(r,\theta)} I(x,y); \tag{3}$$
$$L(r,\theta) = \{(x,y)\,|\, r = x\cos\theta + y\sin\theta\}$$

where $L(r,\theta)$ represents all the image points that lie on a particular line (r,θ).

If we view the time-frequency spectral space of a speech signal as an image plane, it is expected that the HT can help to extract more salient features that would not be possible with short-time spectral analysis only. There are a number of ways that the HT extracted line parameters $\{\hat{r},\hat{\theta}, Acc(\hat{r},\hat{\theta})\}$ can be incorporated into ASR front-end processing. In this work, we investigate the possibility of using the maximum line accumulation $Acc(\hat{r},\hat{\theta})$ to derive a masking threshold for improving noisy speech recognition.

Masking is an important aspect of human audition in which the presence of one sound can raise the hearing threshold of another sound. Masking is particularly relevant to robust ASR front-end processing as it can help to discount the effects of noise on a speech signal. There are two types of masking in the literature, namely, frequency masking and temporal masking [7]. Frequency masking emphasizes the formant regions of a speech signal in the spectral domain, while temporal masking helps to filter out those signal components which change too slowly or too rapidly. In this work, the novel use of HT to derive a masking threshold for simultaneous frequency and temporal masking is investigated.

With the incorporation of HT based masking (HTM) and the addition of time index (t) to the notation, the log Mel-filterbank output of a frame of speech signal at time t is then given by:

$$m_j(t) = \log_e[S_j(t) + \lambda\Phi(t)];$$
$$\Phi(t) = \frac{\sum_{(x,y)\in L(\hat{r},\hat{\theta})} S_y(t - T_w + x)}{n}; \tag{4}$$
$$1 \le x \le T_w; \quad 1 \le y \le M$$

where λ is a constant to control the degree of masking, n is the number of points on the extracted line and T_w is the width of an image formed by the past T_w frames of the Mel-filterbank outputs.

Once a frame of log Mel-filterbank outputs has been masked, discrete cosine transform (DCT) can then be applied to the masked outputs to obtain the corresponding MFCCs.

2.2 Cumulative Distribution Mapping

The cumulative distribution mapping (CDM) method described here is based on the use of histogram equalization (HE) originally developed for improving the contrast of an image [8]. The use of the HE method for compensating handset mismatch in front-end processing of speech can also be found in [9]. The details of our implementation of the CDM can be found in [10]. The main idea of this method is to map the distribution of a time sequence of noisy speech features into a target distribution with a pre-defined probability density function (PDF). In our case, it is assumed that for a given feature value v_o, the mapping is derived from:

$$\int_{v=-\infty}^{v_o} f(v)dv = \int_{z=-\infty}^{z_o} h(z)dz \text{ ; or } F_v(v_o) = F_z(z_o) \tag{5}$$

where $F_v(v)$ is the corresponding cumulative distribution function (CDF) of a given set of noisy speech features and $F_z(z)$ is the target CDF, $f(v)$ and $h(z)$ are the respective PDFs.

From equation (5), the required mapping would be:

$$z_o = F_z^{-1}[F_v(v_o)] \quad (6)$$

Typically, $h(z)$ is assumed to be a Gaussian with zero mean and unity variance. In the experiments, CDM is applied only to the static feature vector which consists of 13 MFCCs ($C_0 \sim C_{12}$) and each cepstral coefficient is normalized individually on a per-utterance basis.

3. EXPERIMENTAL RESULTS

The proposed front-end has been evaluated on the Aurora II database [11] with various configurations. This database contains noisy connected digits (spoken by American adults), which were created by adding various types of noises at different signal-to-noise ratios (SNR) to the original clean (i.e. high SNR) utterances. There are three test sets in the database. Test set *A* contains speech utterances with "subway", "babble", "car" and "exhibition" types of noises, while test set *B* contains speech with "restaurant", "street", "airport" and "train station" types of noises. Test set *C* contains only "subway" and "street" types of noises but there are channel distortions as well. Each of the test sets *A* and *B* has about 28K utterances and the test set *C* has about half of that number.

3.1 Experimental Setup

All pre-processing and Mel-filtering of speech signals followed the ETSI standard MFCC front-end, except that C_0 was used instead of log-energy. There were 23 Mel-filterbanks ($M = 23$) and the static feature vector of our front-end consisted of 13 MFCCs ($C_0 \sim C_{12}$). The static feature vector after noise compensation was appended with its corresponding 1st-order and 2nd-order time derivatives to form a resultant vector with 39 coefficients for speech recognition at the backend. Each recognition model was represented by a continuous density hidden Markov model (HMM) with left-to-right configuration. Digit models had 16 states with 3 Gaussians per state, while the noise model had 3 states with 6 Gaussians per state. All HMMs were trained from a set of 8440 clean utterances which is not included in the test sets.

3.2 Results and Discussion

The official Aurora evaluation framework [11] was followed in that average recognition accuracy for each test set is calculated from the recognition results for those test data with SNRs from 0 dB to 20dB. When the ETSI standard MFCC front-end was used, the average digit accuracy for the test set *A* was found to be 61.34%. If the log-energy (logE) in the ETSI MFCC front-end was replaced with C_0, the average digit accuracy for the same test set was degraded to 58.89%. Although the use of C_0 was found to be less robust than the use of log-energy when there is no noise compensation, it was used in the experiments as it can provide better accuracy when used with CDM. With only the incorporation of HTM ($\lambda = 0.05$, $T_w = 7$ and without CDM) into our front-end, the average digit accuracy was improved to 64.98% from 58.89% [12]. As a baseline here, our front-end with CDM only (without HTM) achieved an average digit accuracy of 81.67% for the test set *A*.

When both HTM and CDM were incorporated into the front-end, some better recognition results for the test set *A* as shown in Table 1 were obtained. This table also summarizes the effects of varying the degree of HT masking (λ) and the image width (T_w) on the recognition accuracy. As observed from the table, $\lambda = 0.05$ and $T_w = 7$ provided the best recognition accuracy at 83.67%. This accuracy represents a relative error rate reduction of 57.8% when compared with the accuracy of the ETSI front-end (61.34%). Moreover, it can be observed that the use of HTM together with CDM provided additional improvement in accuracy over the use of CDM only (83.67% vs. 81.67%). From the results, it seems that the recognition accuracy is less affected by the setting of the image width, but more sensitive to the degree of masking.

Table 1 - Average digit accuracies (%) for Aurora test set *A*, proposed front-end with various settings for HTM

Degree of Masking (λ)	Image Width (T_w)		
	5	7	9
0.5	79.83	80.23	80.34
0.1	82.64	82.80	82.66
0.07	83.09	83.30	83.21
0.05	83.15	**83.67**	83.46
0.03	83.28	83.26	83.57
0.01	82.57	82.53	83.08

The setting $\lambda = 0.05$ and $T_w = 7$ was used to perform further evaluation for the test sets *B* and *C*, and the results are summarized as shown in Table 2. From the table, it can be observed that the proposed front-end outperforms the ETSI standard MFCC front-end for all the three test sets. The proposed front-end is found to achieve higher improvement in accuracy for the test set *B*, but marginally lower improvement for the test set *C*, probably due to the inclusion of channel distortions in the test set *C*. Nevertheless, the average accuracy across the test sets obtained by the proposed front-end represents a relative error rate reduction of 57.1% when compared with that of the ETSI front-end (83.31% vs. 61.08%).

Table 2 - Average digit accuracies (%) for Aurora test sets, proposed front-end compared with ETSI standard MFCC front-end

Front-end	Test *A*	Test *B*	Test *C*	Avg.
ETSI (logE)	61.34	55.75	66.14	61.08
Proposed*	83.67	85.25	81.00	83.31

* $\lambda = 0.05$ and $T_w = 7$

To get an insight on how the combined HTM/CDM front-end is performing in different noise conditions, a break-down of the recognition results according to individual SNR levels and averaged across all three test sets for the front-end

configuration $\lambda = 0.05$ and $T_w = 7$ is shown in Figure 3. From the figure, it can be observed that the proposed front-end achieves better recognition accuracy than that of the ETSI standard front-end at every SNR level. At 5dB SNR, the average accuracy obtained by the proposed front-end is about double that of the ETSI front-end.

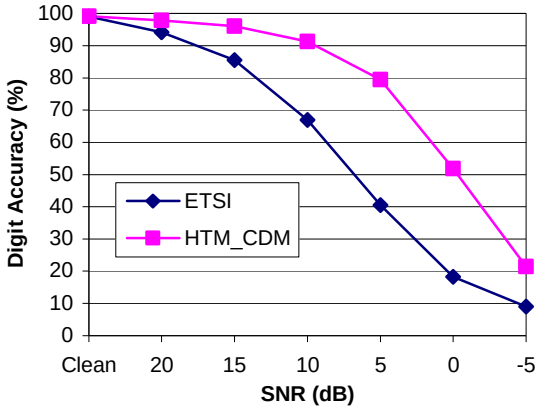

Figure 3 - Recognition results for the Aurora test sets, proposed front-end compared with ETSI standard MFCC front-end by SNR

Figure 4 shows the average recognition results for the test set *A* according to the noise types. It can be observed that overall the biggest improvement is obtained for the "babble" type noisy speech, while the best absolute accuracy is obtained for the "car" type noisy speech, by using the proposed front-end ($\lambda = 0.05$ and $T_w = 7$).

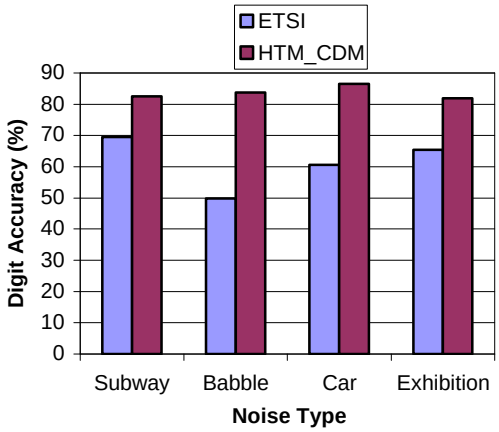

Figure 4 - Recognition results for Aurora test set *A*, proposed front-end compared with ETSI standard MFCC front-end by noise type

Similarly, the recognition results by noise type for the test sets *B* and *C* are also shown in Figure 5. Note that the (C) following the name of a noise type in the figure denotes speech data from the test set *C* which also contains additional channel distortions. Again it can be observed from Figure 5 that the proposed front-end outperforms the ETSI standard front-end for all the different types of noisy speech in these two test sets. The comparatively lower improvement in recognition accuracy for the test set *C* indicates that an additional algorithm for compensating channel distortion more effectively is required.

Overall these two figures demonstrate that the proposed front-end is much more consistent and robust than the ETSI standard MFCC front-end in recognizing different types of noisy speech.

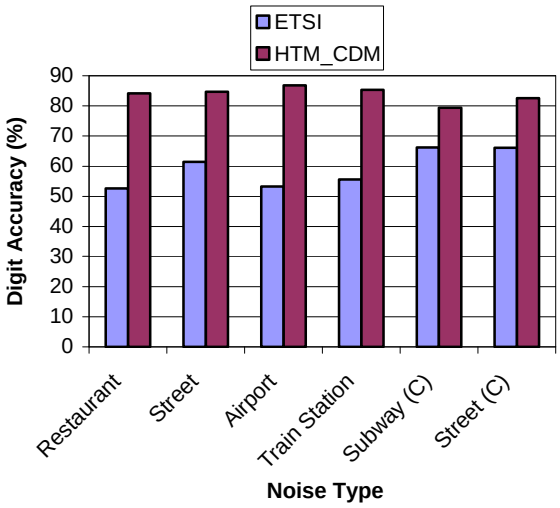

Figure 5 - Recognition results for Aurora test sets *B* and *C*, proposed front-end compared with ETSI standard MFCC front-end by noise type

4. CONCLUSION

A new and noise robust front-end based on the combined incorporation of Hough transform based masking and cumulative distribution mapping has been proposed. Experimental results on the Aurora II speech database have revealed the effectiveness of the novel front-end. The proposed front-end achieves an average digit accuracy of 83.31% for the three Aurora test sets. Future research will focus on the use of other Hough transform extracted line parameters as components of a feature vector, the incorporation of a more effective algorithm for compensating channel distortion, and the use of a different target CDF for the cumulative distribution mapping.

REFERENCES

[1] Huang, C., Wang, H. and Lee, C., "An SNR-Incremental Stochastic Matching Algorithm for Noisy Speech Recognition", *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 8, Nov. 2001, pp. 866-873.

[2] Ephraim, Y., "A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models", *IEEE Trans. Signal Processing*, Vol. 40, No. 4, April 1992, pp. 725-735.

[3] Sankar, A. and Lee, C.H., "A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition", *IEEE Trans. Speech and Audio Processing*, Vol. 4, May 1996, pp. 190–202.

[4] Zhang, Z. and Furui, S., "Piecewise-linear Transformation-based HMM Adaptation for Noisy Speech", *Speech Communication*, Vol. 42, Issue 1, Jan. 2004, pp. 43-58.

[5] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms", *ETSI Standard Document ES 201 108*, April 2000.

[6] Vernon, D., *Machine Vision: Automated Visual Inspection and Robot Vision*, Prentice Hall International (UK), 1991.

[7] Zhu, W. and O'Shaughnessy, D., "Incorporating Frequency Masking Filtering in a Standard MFCC Feature Extraction Algorithm", in *Proc. Int. Conf. on Signal Processing, ICSP'04*, Sept. 2004, pp. 617-620.

[8] Russ, J.C., *The Image Processing Handbook*, CRC Press, 1995.

[9] Dharanipragada, S. and Padmanabhan, M., "A Nonlinear Unsupervised Adaptation Technique for Speech Recognition", in *Proc. Int. Conf. on Spoken Language Processing, ICSLP'00*, Vol. 4, Oct. 2000, pp. 556-559.

[10] Choi, E., "Noise Robust Front-end for ASR using Spectral Subtraction, Spectral Flooring and Cumulative Distribution Mapping", in *Proc. 10th Australian Int. Conf. on Speech Science and Technology*, Dec. 2004, pp. 451-456.

[11] Hirsch, H.G. and Pearce, D., "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noise Conditions", in *Proc. ISCA ITRW ASR2000*, Sept. 2000, pp. 181-188.

[12] Choi, E., "A Noise Robust Front-end for Speech Recognition Using Hough Transform and Cumulative Distribution Mapping", in *Proc. Int. Conf. on Pattern Recognition, ICPR'06*, Aug. 2006, to appear.