ADAPTIVE REGRESSION BASED FRAMEWORK FOR IN-CAR SPEECH RECOGNITION

Weifeng Li♯, Katunobu Itou†, Kazuya Takeda† and Fumitada Itakura‡

Graduate School of Engineering♯, Graduate School of Information Science†, Nagoya University
Faculty of Science and Technology‡, Meijo University
Nagoya, 464–8603 Japan

ABSTRACT

We address issues for improving hands-free speech recognition performance in different car environments using a single distant microphone. In our previous work, we proposed a regression based enhancement method for in-car speech recognition. In this paper, we describe recent improvements and propose a data-driven adaptive regression based speech recognition system, in which both feature enhancement and model compensation are performed. Based on isolated word recognition experiments conducted in 15 real car environments, the proposed adaptive regression approach shows an advantage in average relative word error rate (WER) reductions of 52.5% and 14.8%, compared to original noisy speech and ETSI advanced front-end, respectively.

1. INTRODUCTION

The mismatch between training and testing conditions is one of the most challenging and important problems in automatic speech recognition (ASR). This mismatch may be caused by a number of factors, such as background noise, speaker variation, a change in speaking styles, channel effects, and so on. State-of-the-art ASR techniques for removing the mismatch usually fall into two main categories: feature enhancement and model compensation. Feature enhancement algorithms attempt to transform the corrupted feature into an estimate that more closely resembles clean speech, while model compensation methods aim to adapt or transform acoustic models to match the noisy speech feature in a new testing environment. Examples of the feature enhancement methods include spectral subtraction [1], Wiener filter, CDCN [2], and so on. Spectral subtraction was originally proposed in the context of speech enhancement, but it can be used as a preprocessing step for recognition. However, its performance suffers from inaccurate or erroneous noise estimation. CDCN may be somewhat intensive to compute since it depends on the online estimation of the channel and additive noise through an iterative EM approach. The representative methods in the model compensation category include multi-style training, MLLR [3], and Jacobian adaptation. Their main disadvantage is that they require the retraining of a recognizer or adaptation data. On the other hand, most feature enhancement and model compensation methods are accomplished by linear functions such as simple bias removal, affine transformation, linear regression, and so on. It is well known that distortion caused even by additive noise only is highly nonlinear in the log-spectral or cepstral domain.

The use of a neural network allows us to automatically learn the nonlinear mapping functions between the reference and testing environments. Such a network can handle additive noise, reverberation, channel mismatches, and combinations of these. Neural network based feature enhancement has been used in conjunction with a speech recognizer. For example, Sorensen used a multi-layer network for noise reduction in the isolated word recognition under F-16 jet noise [4]. Yuk and Flanagan employed neural networks to perform telephone speech recognition [5]. However, the feature enhancement they implemented was performed in the ceptral domain and the clean features were estimated using the noisy features only. In a previous work, we proposed a regression based enhancement method for in-car speech recognition [6]. In the proposed method, the log mel-filter-bank (MFB) outputs of clean speech are approximated through the nonlinear regression of those obtained from the noisy speech and estimated noise using a multi-layer perceptron (MLP) neural network. Our neural network based feature enhancement method incorporates noise estimation and can be viewed as generalized log spectral subtraction.

In [6], each driving condition was assumed to be known as a prior information. In this paper, we release this prior information and develop a data-driven speech recognition system, where the regression parameters change adaptively for different driving conditions. To further reduce the mismatch between training and testing conditions, we synthesize the training data using the optimal regression parameters, and train multiple HMMs over the synthesized data. We also develop several HMM selection strategies. The devised system results in a universal in-car speech recognition framework including both the feature enhancement and model compensation.

The organization of this paper is as follows: In Section 2, we describe the in-car speech corpus used in this paper. In Section 3, we present the regression based feature enhancement and environment detection algorithms. In Section 4, we present the adaptive regression based speech recognition framework. Section 5 outlines the performance evaluation and Section 6 summarizes this paper.

2. IN-CAR SPEECH DATA AND SPEECH ANALYSIS

The speech data used are from CIAIR in-car speech corpus [7]. Speech signals are captured by a microphone set on the visor position to the driver. The test data includes Japanese 50 word sets under 15 driving conditions (three driving environments $\times$ five in-car states = 15 driving conditions, as listed in Table 1). For each driving condition, 50 words are uttered by each of 18 speakers. The training data for acoustic modeling comprises a total of 7,000 phonetically balanced sentences (3,600 sentences are collected in the idling-normal condition and 3,400 are collected while driving a data collection vehicle (DCV) around streets near Nagoya University (city-normal condition)). 1,000-state triphone Hidden Markov Modes (HMMs) with 32 Gaussian mixtures per state are used for acoustic modeling.

This work is partially supported by a Grant-in-Aid for Scientific Research (A) (15200014).

Speech signals are digitized into 16 bits at a sampling frequency of 16 kHz. For spectral analysis, a 24-channel MFB analysis is performed on 25-millisecond windowed speech with a frame shift of 10 milliseconds. Spectral components lower than 250 Hz are filtered out to compensate for the spectrum of engine noise, which is concentrated in the lower frequency region. The feature vector used for speech recognition is a 25-dimensional vector (12 CMN-MFCC + 12 Δ CMN-MFCC + Δ log energy).

3. ALGORITHMS

3.1. Regression based feature enhancement

Let $S(m,l)$, $\hat{N}(m,l)$ and $X(m,l)$ denote the log mel-filter-bank (MFB) outputs obtained from the reference clean speech[1], noise and the observed speech signals. m and l denote filter bank and frame indexes, respectively. The hat above N denotes the estimated version. The idea of the regression based enhancement is to approximate $S(m,l)$ with the combination of $X(m,l)$ and $\hat{N}(m,l)$, as shown in Fig. 1. In particular, we estimate $\hat{S}(m,l)$ by applying multi-layer perceptron (MLP) regression method, where a network with one hidden layer composed of 8 neurons is used, i.e.:

$$\begin{aligned}\hat{S}(m,l) &= f(X(m,l),\hat{N}(m,l)) \\ &= b_m + \\ &\sum_{p=1}^{8}\left(w_{m,p}\tanh\left(b_{m,p}+w_{m,p}^{(x)}X(m,l)+w_{m,p}^{(n)}\hat{N}(m,l)\right)\right),\end{aligned}$$

where $\tanh(\cdot)$ is the tangent hyperbolic activation function. The parameters $\Theta = \{b_m, w_{m,p}, w_{m,p}^{(x)}, w_{m,p}^{(n)}, b_{m,p}\}$ are found by minimizing the mean squared error:

$$\mathcal{E}(m) = \sum_{l=1}^{L}[S(m,l) - \hat{S}(m,l)]^2, \qquad (1)$$

through the back-propagation algorithm [8]. Here, L denotes the number of training examples (frames).

Although neural networks have been employed for feature compensation (e.g. [4] [5]) with stereo data, our method incorporates noise estimation and can be viewed as generalized log spectral subtraction. In this paper, the two-stage noise spectra estimator proposed in [6] is used for noise estimation. Based on our previous studies, the incorporation of the noise information offers a benefit of 3% absolute improvement in recognition accuracies, compared to that using the noisy features only.

[1]Speech collected with a close-talking microphone (with a headset) is used for reference clean speech.

Table 1. 15 driving conditions (3 driving environments $\times$ 5 in-car states)

driving environment	idling city driving expressway driving
in-car state	normal CD player on air-conditioner (AC) on at low level air-conditioner (AC) on at high level window (near driver) open

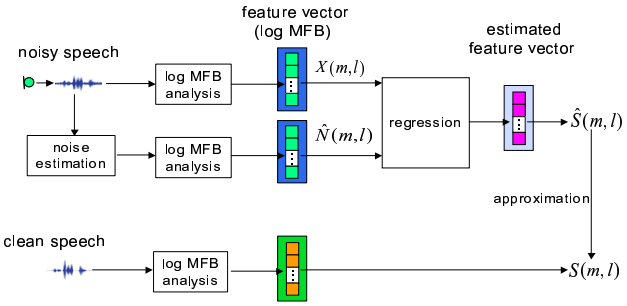

Fig. 1. Concept of regression-based feature enhancement.

3.2. Driving environment detection

In this subsection, we discriminate in-car environments by using the information of noise signals. Noise classification is a nontrivial task in our studies since the difference among driving conditions is not significant. An important step is feature selection. In our studies, Mel-frequency cepstral coefficients (MFCC) were selected because of their good discriminating ability, even in audio classification (e.g. [9] [10]). The MFCC features are extracted frame by frame from non-speech signals (preceding the utterance by 200 ms, i.e., 20 frames), their means in one noisy signal are computed, and they are then concatenated into a feature vector:

$$\mathbf{R} = [\mu_1, \ldots, \mu_{12}, \mu_e], \qquad (2)$$

where μ_i and μ_e denote the means of i-order MFCC and log energy, respectively. All of the 13 elements in $\mathbf{R}$ are normalized so that their mean and variance across the elements are 0 and 1.0, respectively. Prototypes of the noise clusters are obtained by applying the *K-means-clustering* algorithm to the feature vectors extracted from the training set of noise signals. In our experiments, the non-speech signals by 12 speakers are used to cluster the noise conditions, and those by another six speakers are used for testing, as shown in Fig. 2.

4. ADAPTIVE REGRESSION BASED SPEECH RECOGNITION

4.1. Regression-based HMM training

In our previous work [11], we generated the enhanced speech signals, by performing the regression in the log spectral domain (for each frequency bin). Though few "musical tone" artifacts were found in the regression-enhanced signals compared to those obtained using spectral subtraction based methods, some residual noise still existed in the regression-enhanced signals. We believe there will exit a mismatch between training and testing conditions, if we use HMM trained over clean data to test regression-enhanced speech. Therefore, we adopt the K sets of optimal weights obtained from each clustered group to generate 7,000-sentence training data, i.e., we simulated 7,000 $\times$ K sentences based on K clustered noise environments. Next, K HMMs are trained over each of the simulated 7,000-sentence training data. On the other hand, because multi-style training has been shown to be effective for improving the ASR performance [12], a universal HMM is also trained over all the simulated 7,000 $\times$ K sentences, as shown in Fig. 2.

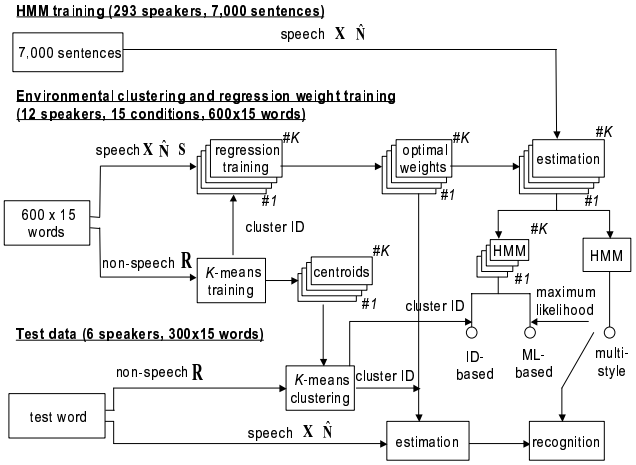

Fig. 2. Diagram of adaptive regression based speech recognition. $\mathbf{X}$, $\hat{\mathbf{N}}$, and $\mathbf{S}$ denote the log MFB outputs obtained from observed noisy speech, estimated noise, and reference clean speech, respectively. $\mathbf{R}$ denotes the vector representation of driving environment using Eq. (2).

4.2. Adaptive recognition of an input signal

The recognition of an input signal consists of two phases: the feature enhancement phase and the HMM selection phase. In the feature enhancement phase, for unknown input speech, we find a corresponding noise group through the non-speech segments and perform the estimation with the optimal weights for the noise cluster, i.e., the log MFB outputs of clean speech can be estimated by

$$\hat{\mathbf{S}} = f_k(\mathbf{X}, \hat{\mathbf{N}}) \tag{3}$$

where $\mathbf{X}$ and $\hat{\mathbf{N}}$ indicate the log MFB vector obtained from noisy speech and estimated noise respectively. $f_k(\cdot)$ corresponds to the nonlinear mapping function in Section 3.1, where the cluster ID k is specified by minimizing the Euclidian distance between $\mathbf{R}$ and the centroid vectors.

In the HMM selection phase, beside a universal HMM obtained using multi-style training, an HMM is selected from K HMMs based on the following two strategies:

1. ID-based
   This strategy tries to select an HMM that is trained with the simulated training data close to the test noise environment, i.e.,

$$\hat{H}(x) = \sum_{k=1}^{K} I(D(x), D(H_k)) H_k \tag{4}$$

   where

$$I(D(x), D(H_k)) = \begin{cases} 1, & \text{if } D(x) = D(H_k) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

   and $D(x) = D(H_k)$ means that the cluster ID of an input signal x is identical to that of kth HMM H_k.

2. maximum likelihood (ML) based
   This strategy tries to select the HMM that outputs maximum likelihood, i.e.,

$$\hat{H}(x) = \arg\max_{H} \{P(x|H_1), \ldots, P(x|H_K)\} \tag{6}$$

   where $P(x|H_k)$ indicates the log likelihood of an input signal x by using kth HMM H_k.

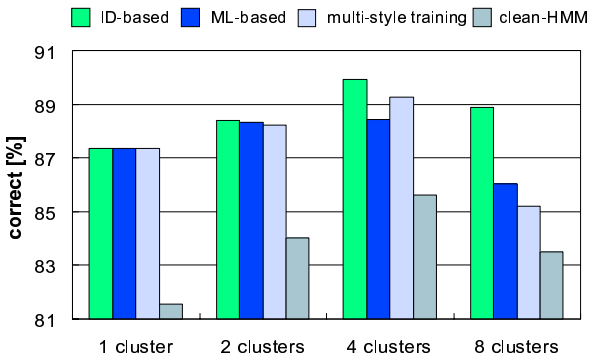

Fig. 3. Recognition performance for different clusters using adaptive regression methods (averaged over 15 driving conditions).

4.3. Analysis of the proposed frmaework

There are some common points in SPLICE [13] and our feature enhancement phase. Both of them are stereo-based and consist of two steps: finding the optimal "codeword" and performing codeword-dependent compensation (see Eq. (3)). However, the proposed enhancement method does not need any Gaussian assumption required in SPLICE and turns out to be a nonlinear compensation. Regression-based HMM training and HMM selection can be viewed as a kind of nonlinear model compensation, which can incorporate the information of the testing environments. A combination of feature enhancement and HMM selection results in a universal speech recognition framework where both the noisy features and acoustic models are compensated.

5. PERFORMANCE EVALUATION

Figure 3 shows the word recognition accuracies for different numbers of clusters using adaptive regression methods. It is found that the recognition performance is improved significantly by using adaptive regression methods compared to those of "clean-HMM", which is trained over the speech at the close-talking microphone. As the number of clusters increases up to four, the recognition accuracies increase consistently due to there being more noise information available, however too many clusters (e.g., eight or above) yields a degradation of the recognition performance. Although the three adaptive regression based recognition systems perform almost identically in the two-cluster case, "ID-based" yields a more stable recognition performance across the numbers of clusters, and the best recognition performance is achieved with four clusters.

For comparison, we also performed recognition experiments based on the originally observed noisy speech ("original"), a MAP speech amplitude estimator ("MAP") [14], ETSI advanced front-end [15], and an adaptive beamformer ("ABF"; Four linearly spaced microphones with an inter-element spacing of 5 cm at the visor position are used.). The acoustic models used for "MAP", ETSI advanced front-end and adaptive beamforming were trained over the training data they processed. Figure 4 shows the recognition performance averaged over the 15 driving conditions. "proposed" cites the best recognition performance achieved in Fig. 3. It is found that all the enhancement methods outperform the original noisy speech. ETSI advanced front-end yields higher recognition accuracy than MAP. The proposed method significantly outperforms ETSI advanced front-end and even performs better than adaptive beamforming, which uses as many as four microphones. Recalling Fig. 3, it is found

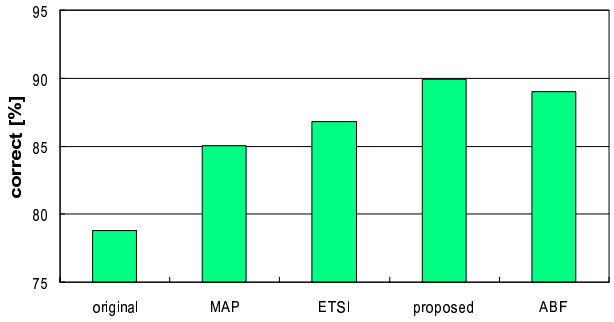

Fig. 4. Recognition performance of different speech enhancement methods (averaged over 15 driving conditions).

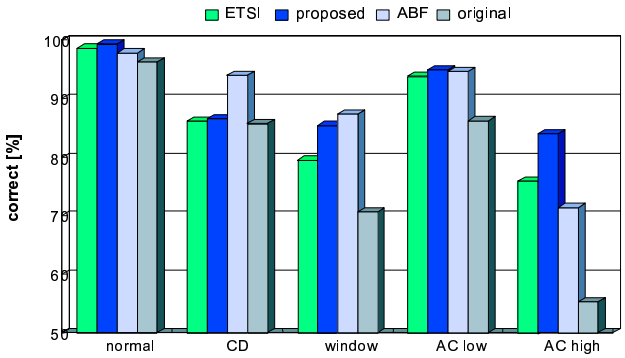

Fig. 5. Recognition performance for five in-car states shown in Table 1.

that even one cluster using the proposed method outperforms ETSI advanced front-end. This clearly demonstrates the superiority of the adaptive regression method.

We also investigated the recognition performance averaged over five in-car states as shown in Fig. 5. It is found that the adaptive regression method outperforms ETSI advanced front-end in all the five in-car states, especially when AC is on at high level and when the window near the driver is open. Adaptive beamforming is very effective when the CD player is on and when the window near the driver is open. This suggests that adaptive beamforming with multiple microphones can suppress the noise coming from undesired directions quite well due to its spatial filtering capability. However, in the remaining three in-car states (diffuse noise cases), it does not work as well as the adaptive regression method.

6. SUMMARY

In this paper, we have proposed a data-driven adaptive regression based speech recognition system, which includes the driving environmental detection, the regression based feature enhancement, and the HMM selection. The devised system turns out to be a universal speech recognition framework that performs both feature enhancement and model compensation. The superiority of the proposed system was demonstrated by a significant improvement in recognition performance in the isolated word recognition experiments conducted in 15 real car environments.

7. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol.ASSP-27, no.2, pp.113-120, 1979.

[2] A. Acero, "Acoustical and environmental robustness in automatic speech recognition," Ph.D. Thesis, Carnegie Mellon University, 1990.

[3] C. Leggetter and P.C.Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, 9:171-186, 1995.

[4] Helge B.D. Sorensen, "A cepstral noise reduction multi-layer neurla network," in *Proc. IEEE ICASSP*, pp. 933-936, 1991.

[5] D. Yuk and J. Flanagan, "Telephone speech recognition using neural networks and hidden markov models,", in *Proc. IEEE ICASSP*, pp. 153-156, 1999.

[6] W. Li, K. Itou, K. Takeda and F. Itakura, "Two-stage noise spectra estimation and regression based in-car speech recognition using single distant microphone," in *Proc. IEEE ICASSP*, pp. I-533-536, 2005.

[7] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura, and Y. Inagaki, "Construction of speech corpus in moving car environment," in *Proc. ICSLP*, pp.362-365, 2000.

[8] S. Haykin, *Neural Networks - A Comprehensive Foundation*, Prentice-Hall, 1999.

[9] M.J. Carey, E.S. Parris, and H.L. Thomas, "A comparison of features for speech, music discrimination," in *Proc. IEEE ICASSP*. pp. 149-152, 1999.

[10] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proc. IEEE ICASSP*, pp.1941-1944, 2002.

[11] W. Li, K. Itou, K. Takeda and F. Itakura, "Subjective and objective quality assessment of regression-enhanced speech in real car environments," in *Proc. Interspeech'2005-Eurospeech*, pp. 2093-2096.

[12] R.P.Lipmann, E.A.Martin and D.B.Paul, "Multi-style training for robust isolated word speech recognition," in *Proc. IEEE ICASSP*, pp.705-708, 1987.

[13] L. Deng, A. Acero, M. Plumpe and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. ICSLP*, pp. 806-809, 2000.

[14] P.J. Wolfe and S.J. Godsill, "Effective alternatives to the ephraim and malah suppression rule for audio speech enhancement," *EURASIP Journal on Applied Signal Processing*, vol.2003, no.10, pp.1043-1051, 2003.

[15] "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm," ETSI ES 202 050 v1.1.1, 2002.