Robust Automatic Speech Recognition for Accented Mandarin in Car Environments

Pei Ding, Lei He, Xiang Yan and Jie Hao

Toshiba (China) Research and Development Center, Beijing, China
{dingpei, helei, yanxiang, haojie}@rdc.toshiba.com.cn

Abstract

This paper addresses the issues of robust automatic speech recognition (ASR) for accented Mandarin in car environments. A robust front-end is proposed, which adopts a Minimum Mean-Square Error (MMSE) estimator to suppress the background noise in frequency domain, and then implements spectrum smoothing both in time and frequency index to compensate those spectrum components distorted by the noise over-reduction. In the context of Mandarin speech recognition, a special adverse factor is the diversification of Chinese dialects, i.e. the pronunciation difference among dialects decreases the recognition performance if the acoustic models are trained with an unmatched accented database. We propose to train the models with multiple accented Mandarin databases to solve this problem. Evaluation results of isolated phrase recognition show that the proposed front-end can obtained the average error rate reduction (ERR) of 58.3% and 9.7% for artificial car noisy speech and real in-car speech respectively, when compared with the baseline in which no noise compensation technology is used. The efficiency of the proposed model training scheme is also proved in the experiments.

Index Terms: robust speech recognition, in-car speech, MMSE enhancement, spectrum smoothing, accented Mandarin

1. Introduction

In recent years an important application of ASR technologies is to act as a voice-activated human-machine interface in car navigation systems. These embedded ASR modules provide a safe and convenient input method, and usually make good balances between usable functionality and system complexity. Consequently, such devices become very popular and many kinds of mass-produced cars have been equipped.

Among the difficulties in such in-car speech recognition tasks, the most critical problem is to cope with the background noise, which is incurred by mechanical oscillation of engine, friction between the road and tires, blowing air outside the car, and etc. Noise robustness is the common challenge of ASR systems, and many approaches have been proposed for this issue [1]. Some methods [2][3][4][5] aim at designing a robust front-end in which the interfering noise is removed from the speech or the acoustic feature is inherently less distorted by noise. Other methods [6][7][8] are concentrated on model adaptation technologies which decrease the mismatch between noisy speech features and the pre-trained acoustic models. Generally, model adaptation methods are superior to those that extract robust features [9], but their major disadvantage is that they usually cause huge computation load. Besides, the less dependency between the front-end and the recognizer can effectively reduce the complexity of ASR systems. In this paper, we propose a robust front-end, in which MMSE estimation algorithm [10] is used to suppress the noise in frequency domain. Compared to other conventional speech enhancement algorithms, such as spectral subtraction (SS) [2], the MMSE estimation is more efficient in minimizing both the residual noise and speech distortion. In speech enhancement, some spectrum components at very low signal-to-noise ratios (SNR) tend to be floored by meaningless threshold in Mel-scaled filter binning stage because of the noise over-reduction. Even not floored, these spectrum components are prone to aggressively degrade the recognition performance. We propose to smooth the spectrum both in time and frequency index with arithmetic sequence weights. Thus, those unreliable spectrum components will be fed with speech energy from neighbors with high local SNRs, and the recognition rate can be efficiently improved.

In the context of Mandarin speech recognition, an inevitable problem is the diversification of Chinese dialects. The dialectal pronunciation characteristic will affect the style of uttering Mandarin speech and cause the phonetic and acoustic confusion [11]. If the speaker has a regional accent different from the standard Mandarin on which the acoustic models are trained, the recognition performance will be degraded. Representative methods for this issue include speaker and model adaptation [12][13], accent detection and model selection [14], and pronunciation modeling [13]. In this paper, we propose an acoustic model training scheme that uses multiple accented Mandarin databases, and confirm its efficiency in isolated phrase recognition task.

The rest of the paper is organized as follows. Section 2 describes the proposed noise robust front-end. Section 3 analyses the problem of accented Mandarin speech recognition and introduce our model training scheme for this issue. Section 4 and section 5 describes the experiments in details. Finally, section 6 concludes the paper.

2. Noise robust front-end

2.1. MMSE estimation algorithm

Ephraim and Malah proposed a short-time spectral amplitude (STSA) estimation algorithm based on a MMSE criterion to enhance the noise corrupted speech [10]. One advantage is that MMSE estimation algorithm can efficiently suppress the background noise while at the expense of very few speech distortions. Another property of this method is that it can eliminate the residual "musical noise".

We assume that the noise is additive and independent of the clean speech, and after fast Fourier transform (FFT) analysis of windowed speech frames each spectral component is statistical independent and corresponds to a narrow-band Guassian stochastic process. Let $A(k, n)$, $D(k, n)$ and $R(k, n)$ denote the *kth* spectral component of the *nth* frame of speech, noise, and the observed signals respectively, the estimation of $A(k,n)$ is given as

$$\hat{A}(k,n) = \frac{1}{2}\sqrt{\frac{\pi\xi_k}{\gamma_k(1+\xi_k)}}M(-0.5;1;-\frac{\gamma_k\xi_k}{1+\xi_k})R(k,n) \qquad (1)$$

where $M()$ is the confluent hyper-geometric function, the *a priori* SNR ξ_k and the *a posterior* SNR γ_k are defined as :

$$\xi_k \triangleq \frac{E(|A(k,n)|^2)}{E(|D(k,n)|^2)}; \quad \gamma_k \triangleq \frac{|R(k,n)|^2}{E(|D(k,n)|^2)}. \qquad (2)$$

In practice, we use a voice activity detection (VAD) based noise estimation method and substitute the estimation of clean speech by the enhanced spectra of previous frame.

2.2. Spectrum smoothing after MMSE enhancement

The MMSE enhancement algorithm can be interpreted as it suppresses or emphasizes the spectrum components according to their local SNRs. The speech signals in those components at very low SNRs will be seriously distorted due to the noise over-reduction.

Our proposed front-end is based on the framework of cepstral feature extraction, in which a threshold is usually essential to eliminate the sensitivity of logarithmic transform to very small outputs of the Mel-scaled filters. Thus, after speech enhancement, those low SNR spectrum components tend to be floored by a meaningless threshold in Mel-scaled filter binning stage, which causes the mismatch between the features and the acoustic models. Even over the thresholds, the low SNR components are also prone to aggressively degrade the recognition performance.

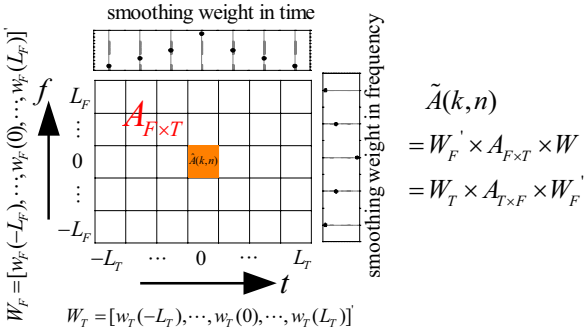


Figure 1: Spectrum smoothing in time and frequency index

In order to compensate the spectrum components distorted by noise over-reduction, we propose to smooth the spectrum both in time and frequency index with symmetrical normalized arithmetical sequence weights. The unreliable spectrum component will be filled with speech energy from neighboring bins whose local SNRs are high and avoid being floored in binning stage, consequently. Thus, the implementation of MMSE enhancement is tamed towards ASR tasks and the recognition performance is efficiently improved further.

At frame *n* and frequency band *k*, the smoothed spectrum component $\tilde{A}(k,n)$ is obtained as follows:

$$\begin{aligned}\tilde{A}(k,n) &= \sum_{i=-LF}^{i=LF}\sum_{j=-LT}^{j=LT} w_F(i)\times w_T(j)\times \hat{A}(k+i,n+j) \\ &\triangleq W_F' \times A_{F\times T} \times W_T \\ &= W_T \times A_{T\times F} \times W_F'\end{aligned}, \qquad (3)$$

where $w_F(i)$ is the arithmetic sequence weight in the frequency index with smoothing length $F = 2\times L_F + 1$:

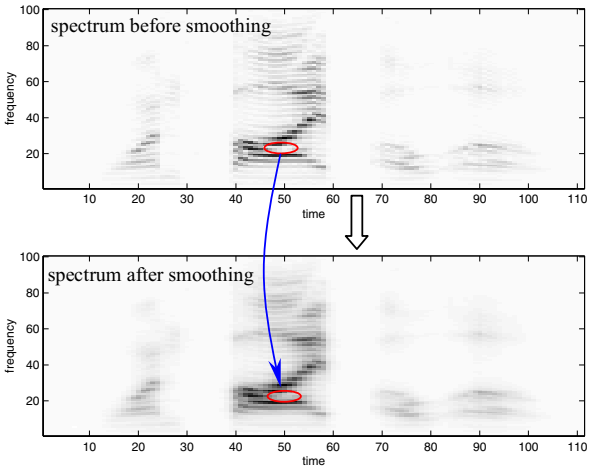


Figure 2: Spectrum examples before and after smoothing

$$w_F(i) = w_F(-i) = \frac{(1-w_F(0))(L_F+1-i)}{L_F(L_F+1)}, 1 \le i \le L_F, \qquad (4)$$

$W_F = [w_F(-L_F), \cdots, w_F(0), \cdots, w_F(L_F)]'$ and $w_F(0)$ is the weight of current frequency bin. $w_T(j)$ and W_T are the smoothing weights in time index and have the similar definitions. The matrix $A_{F\times T}$ corresponds to the spectrum block that is used for smoothing. As illustrated in Fig.1, in Eq.(3) the expression in matrix multiplication style indicates that we can firstly smooth the spectrum in frequency index and then in time index, or equivalently reverse the order. Fig.2 gives the spectrum examples before and after smoothing, and it is very obvious that the low SNR components effectively acquire the speech energy from neighboring bins.

3. Robust ASR for accented Mandarin

A major difficulty in Mandarin speech recognition is to cope with various accented pronunciation. In China, there are seven major dialects and each has a particular pronunciation property. For example, /zh/ is usually pronounced as /z/ by speakers in Wu dialect regions, whose representative city is Shanghai. Such phenomena in accented Mandarin cause both phonetic and acoustic confusions. If the ASR system is trained on a standard or a certain dialectal Mandarin database, it will fail to perform well when the speaker has a different regional accent compared to recognizing the matched speech. In real applications diversification of Chinese dialects is unavoidable, thus robustness to accented Mandarin is a critical issue in designing a universal ASR system for all kinds of accents.

In this paper we propose to train the acoustic models by multiple accented Mandarin speech databases. With this training scheme, the acoustic models are capable of covering statistical characteristics of possible accented pronunciations under moderate model size. Evaluations prove its efficiency in isolated phrase recognition task, which is commonly adopted in the scenario of in-car speech recognition. Besides, using a uniform acoustic model trained on multiple dialectal databases has the advantage to make the ASR system flexible and reduce its complexity.

4. Experiment setup

4.1. Front-end configurations

In the experiments, the speech data are sampled at 11025Hz and 16 bits quantization. The frame length and window shift are 23.2ms and 11.6ms, respectively. In spectra processing, after MMSE speech enhancement and spectrum smoothing, 24 triangle Mel-scaled filters are applied to combine the frequency components in each bank, and the outputs are compressed by logarithmic function. Then the Discrete cosine transform (DCT) decorrelation is performed on the log-spectrum. The final acoustic feature of each frame is a 33 dimensional vector consisting of 11 Mel frequency cepstral coefficients (MFCC) and their first and second order derivatives.

4.2. Acoustic models

For Mandarin speech recognition on isolated phrase task, we adopt the model structure with moderate complexity, in which each Mandarin syllable is modeled by a right-context-dependent INITIAL (bi-phone) plus a toneless FINAL (mono-phone). Totally, there are 101 bi-phone, 38 mono-phone and one silence hidden Markov models (HMM). Each model consists of 3 emitting left-to-right states with 16 Gaussian mixtures.

4.3. Databases

Three accented Mandarin speech databases, denoted as ACM1, ACM2 and ACM3, are used for the evaluations, each of which was collected in the representative city of the corresponding dialectal region. The three are the major dialects in China and the accents are quite different from each other. The speakers are native and the speech data are recorded in quiet environments. Each database includes 50000 utterances in training set and 2000 in testing set. In recognition tasks, the vocabulary includes 500 isolated phrases.

To improve the robustness of ASR system we use an immunity learning scheme [15] in which the acoustic models are trained in simulated noisy environments by artificially adding car noises to clean training utterances at different SNRs. There are 12 kinds of car noises in the experiments, which are the combinations of the following three conditions:

(1) Speed (km/h): 40, 60 and 100
(2) Road type: “asf” (asphalt), “tun”(tunnel) and “con” (concrete)
(3) Air-conditioner state: on/off.

We also generate artificial in-car noisy test speech for evaluation.

4.4. Real in-car evaluation speech data

To evaluate the proposed ASR system in realistic scenarios, in-car Mandarin speech data are collected from native speakers in the same three cities mentioned above. The speech is recorded in the car cabinet through a distant microphone placed in the roof lamp under idling or driving (speed is around 100km/h) conditions.

5. Evaluation results and analysis

5.1. Accented Mandarin recognition experiments

To improve the robustness for different accented Mandarin speech recognition, we propose to train the acoustic models on multiple accented Mandarin databases, i.e. the three training sets from ACM1, ACM2 and ACM3 are merged into one that is denoted as ACM3in1. We also train the models with single accented database for comparison.

Table 1. Experimental results (WER, %) of clean accented Mandarin speech recognition task.

Training Scheme	Accented Testing Set			Ave.
	ACM1	ACM2	ACM3	
ACM1	1.30	5.20	2.90	3.13
ACM2	2.90	1.15	2.70	2.25
ACM3	2.65	2.65	1.70	2.33
ACM3in1	1.55	1.85	2.20	1.87

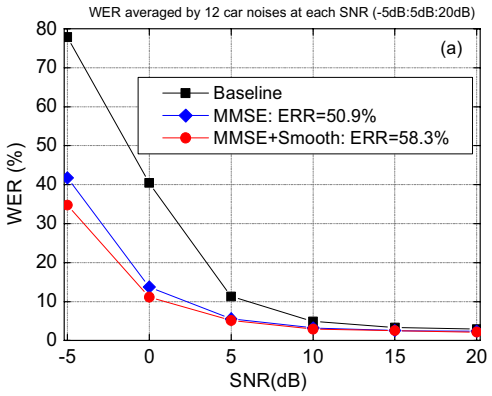
Table 1 shows the word error rate (WER) results in clean accented Mandarin recognition experiments. From the results we can find that if the acoustic models are trained on a certain accented Mandarin database, the recognition performance is very high to deal with the same accented speech, but dramatically degrades in the cross testing with another dialectal accent. For example, in training scheme ACM1, the WER for matched testing set is 1.30% and drastically drops to 5.20% and 2.90% when dealing with the speech from ACM2 and ACM3, respectively. The proposed training scheme, ACM3in1, shows the robustness to the variation of dialectal pronunciation and provides consistent satisfying performance for each accented Mandarin testing set. Compared to the ACM1, ACM2 and ACM3 scheme, the proposed ACM3in1 scheme achieves the average ERR of 40.3%, 16.9% and 19.7%, respectively.

5.2. Evaluations on artificial in-car noisy speech

Twelve car noises are artificially added to clean speech at different SNRs from –5dB to 20dB with a 5dB step to evaluate the robustness of our proposed front-end. The clean test set consists of 600 utterances from the three accented Mandarin databases and totally there are 72 corresponding noisy versions (12 car noises × 6 SNRs). Fig. 3 shows the experiments results, in which the baseline refers to the standard MFCC without any noise robust technology.

Fig. 3(a) shows the WER averaged by 12 car noises at each SNR. We can observe that the baseline front-end performs well in high SNR conditions, e.g. at 20dB the recognition accuracy is 97.07%. However, interfering noises cause serious mismatch between the features and the models, and below 10dB the baseline performance drops rapidly. If the MMSE speech enhancement algorithm is adopted, the noise will be efficiently suppressed in the spectrum. The front-end applying MMSE enhancement significantly improves the robustness, when compared to the baseline. Spectrum smoothing both in time and frequency index tames the implementation of MMSE speech towards ASR application, and from the results it is very obvious that the spectrum smoothing algorithm further improves the recognition performance of MMSE method. In the experiments, the proposed MMSE-Smooth scheme achieves the average ERR of 58.3% versus the baseline.

Fig. 3(b) gives the WER averaged by the six SNRs, from which the performance difference under each car noise is analyzed. We find that the recognition performance in air-conditioner on and high speech driving conditions is obviously lower than in the opposite conditions. The reason is that ASR performance tends to be degraded more seriously by broadband noises. In such adverse environments mentioned above the dominant noise source is the air friction from the air-conditioner and the wind outside the car, which produces the broadband white-like background noises and consequently causes dramatic performance drop for recognition. The experimental results also show that the proposed front-end can significantly improve the performance in all conditions.



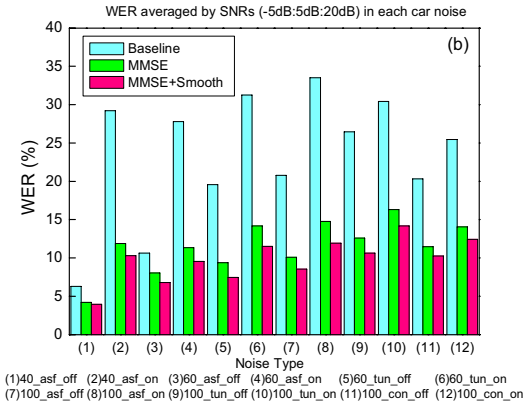


Figure 3. Evaluation results in artificial in-car noisy environment.

5.3. Evaluations on real in-car speech

The proposed front-end is also evaluated on real in-car speech database, as showed in Fig. 4. There are 4795 utterances in idling state test set and 4810 in driving state test set, respectively. From the experimental results, it can be concluded that the proposed method efficiently improves the robustness for real in-car speech recognition task and achieves the average ERR of 9.7% versus the baseline. The ERR here is much lower than that in artificial in-car noisy speech evaluations. The reason is that the background car noises of the in-car test speech are not so aggressive and the typical SNR is about 10dB.

6. Conclusions

This paper presents a robust ASR system for accented Mandarin speech in car environments. In the front-end, a MMSE estimation algorithm is utilized to efficiently suppress the background noises, and then the noise over-reduced spectrum components are compensated by smoothing the enhanced spectrum both in time and frequency index with arithmetic sequence weights. To cope with various dialectal pronunciation styles in Mandarin speech, the acoustic models are trained on multiple accented Mandarin databases. It can be concluded from the evaluation results that our proposed methods can efficiently improve the robustness against both the car noise and accent variation in Mandarin speech.

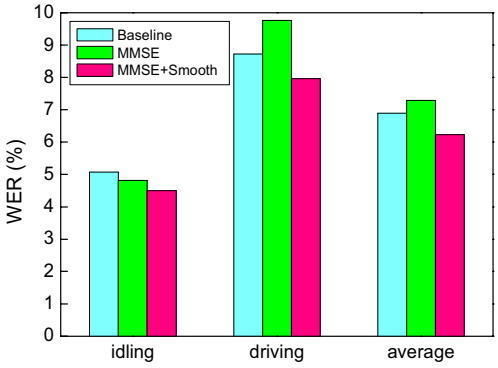


Figure 4. Evaluation results in real in-car speech recognition.

7. References

[1] Y. Gong, “Speech recognition in noisy environments: a survey”, *Speech Communication*, Vol. 16, pp. 261-291, 1995.

[2] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction”, *IEEE Trans. Acoust. Speech and Signal Processing*, Vol. ASSP-27, pp.113-120, 1979.

[3] O. Viikki, D. Bye and K. Laurila, “A recursive feature vector normalization approach for robust speech recognition in noise”, *in Proc. Of ICASSP*, 1998, pp. 733-736.

[4] ETSI Standard, “Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms”, ETSI ES 202 050 v.1.1.1, October 2002.

[5] B. Mak, Y. Tam and Q. Li, “Discriminative auditory features for robust speech recognition”, *in Proc. Of ICASSP*, 2002, pp. 381-384.

[6] M. J. F. Gales and S. J. Young, “Robust continuous speech recognition using parallel model combination”, *IEEE Trans. on SAP*, Vol.4, No. 5, pp. 352-359, 1996.

[7] P. J. Moreno, B. Raj and R. M. Stern, “A vector Taylor series approach for environment-independent speech recogntion”, *in Proc. Of ICASSP*, 1995, pp. 733-736.

[8] H. Shimodaira, N. Sakai, M. Nakai and S. Sagayama, “Jacobian joint adaptation to noise, channel and vocal tract length”, *in Proc. Of ICASSP*, 2002, pp. 197-200.

[9] L. Brayda, L. Rigazio, R. Boman and J. Junqua, “Sensitivity analysis of noise robustness methods”, *in Proc. Of ICASSP*, 2004, pp. 1037-1040.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Trans. Acoustic, Speech, and Signal Processing*, Vol. ASSP-32, pp.1109-1121, 1984.

[11] Y. Liu and P. Fung, “Acoustic and phonetic confusions in accented speech recognition”, *in Proc. of Eurospeech*, 2005, pp.3033-3036.

[12] C. Huang, E. Chang, J. Zhou, and K. Lee, "Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition,” *in Proc. of ICSLP,* 2000, pp. 818–821.

[13] Y. Liu and P. Fung, “State-dependent phonetic tied mixtures with pronunciation modeling for spontaneous speech recognition”, *IEEE Trans. on SAP*, Vol.12, No. 4, pp. 351-364, 2004.

[14] Y. Zheng, R. Sproat, L. Gu, I. Shafran and etc., "Accent detection and speech recognition for Shanghai-accented Mandarin", *in Proc. of Eurospeech*, 2005, pp.217-220.

[15] Y. Takebayashi, H. Tsuboi and H. Kanazawa, “A robust speech recognition system using word-spotting with noise immunity learning”, *in Proc. Of ICASSP*, 1991, pp. 905-908.