A NOISE ROBUST FRONT-END WITH LOW COMPUTATIONAL COST FOR EMBEDDED IN-CAR SPEECH RECOGNITION

Pei Ding, Lei He, Xiang Yan, Rui Zhao, Jie Hao

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

ABSTRACT

This paper proposes a noise robust front-end with low computational cost for embedded in-car speech recognition. The minimum mean-square error (MMSE) estimation algorithm is adopted to suppress the background noise, and in the gain function calculation a suitable piece-wise linear function is used to substitute the traditional Taylor series accumulation method to simplify the computation complexity. After speech enhancement, spectrum smoothing is implemented in both time and frequency index with geometric sequence weights to further compensate the spectral components distorted by noise over-reduction. Experiments on Chinese isolated phrase recognition show that the proposed front-end significantly improves the recognition robustness in car environments while the computational load is extremely reduced. Compared with the ETSI advanced front-end, the average error reduction rate (ERR) of 12.2% and 4.5% is obtained in artificial car noisy speech and real in-car speech, respectively.

Index Terms— speech recognition, acoustic noise, speech enhancement, robustness

1. INTRODUCTION

In recent years an important application of automatic speech recognition (ASR) technologies is to act as a voice-activated human-machine interface in car cabinet for navigation system as well as device control. These embedded ASR modules provide a safe and convenient input method, and usually make good balances between usable functionality and system complexity. Consequently, such hands-free devices attract many attentions 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 ambient acoustic noise, which is incurred by mechanical oscillation of engine, friction between the road and tires, and aerodynamic turbulence, e.g. the air blowing the car body. Noise robustness is the common challenge to ASR systems. Many approaches have been proposed for this issue [1] and can be roughly classified into two categories. Methods in the first category aim at designing a robust front-end in which the acoustic feature is inherently less distorted by noise [2] or the interfering noise is removed by using feature compensation [3] or speech enhancement methods [4][5]. Due to the high non-linearity introduced by mathematic transforms in feature extraction, feature compensation methods usually need stereo data to estimate the statistical models or mapping function parameters [6][7]. Whereas for enhancement methods performed in signal space, the relation between noise and speech can be simplified and the clean speech can be estimated accurately based on some reasonable prior statistical hypothesis [8][9]. Robust methods in the second category concentrate on model adaptation in which the mismatch between noisy speech features and the pre-trained acoustic models is compensated [10][11][12]. Generally, model adaptation methods use more prior statistical information of speech and are superior to those that extract robust features, but their major disadvantage is that they usually cause huge computational load. Besides, the less dependency between the front-end and the recognizer can effectively reduce the complexity of ASR systems.

This paper proposes a robust front-end based on cepstral feature extraction framework, in which MMSE estimation algorithms [8][9] are used to suppress the noise in frequency domain. Compared to other conventional speech enhancement algorithm, such as spectral subtraction [4], the MMSE estimation method is more efficient in minimizing both the residual noise and speech distortion.

In MMSE estimation algorithm, the gain function is calculated by Taylor series accumulation method, which results in the huge computational load and is the weakness of MMSE estimator especially for embedded ASR systems. Storing the pre-calculated function values in a lookup table is a common solution, but the large extra memory cost still restricts the application in such resource-limited situations. In this paper we propose to use a proper piece-wise linear function to substitute the gain function according to the derivative difference and approximation error. Thus, the computational load can be extremely reduced, while the same noise reduction performance is maintained.

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 in both time and frequency indexes with geometric sequence weights. Thus, those unreliable spectrum components will be fed with speech energy from neighboring bins with high local SNRs, and the recognition rate can be efficiently improved.

The rest of the paper is organized as follows. Section 2 describes the MMSE estimation algorithm and the approximation of the gain function. Section 3 introduces the spectrum smoothing algorithm. Section 4 and 5 describe the experiments in details. Finally, section 6 concludes the paper.

2. NOISE REDUCTION ALGORITHM

2.1. MMSE estimation algorithm

Ephraim and Malah proposed the short-time spectral amplitude (STSA) estimation algorithm with a MMSE criterion for the linear-spectra (LinMMSE) [8] or the Log-spectra (LogMMSE) [9]. 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 the residual "musical noise" can be efficiently eliminated. In resent extensive subjective comparison among the representative speech enhancement algorithms [13], the above statistical-model based methods perform the best.

It is assumed in the prior statistical hypothesis that the noise is additive and uncorrelated to 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 Gaussian 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 $A(k,n)$ in LinMMSE algorithm 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 $M(a;c;x)$ is the confluent hyper-geometric function that is calculated by Taylor series accumulation as follows

$$M(a;c;x) = 1+\frac{a}{c}\frac{x}{1!}+\frac{a(a+1)}{c(c+1)}\frac{x^2}{2!}+\cdots=\sum_{r=0}^{\infty}\frac{(a)_r}{(c)_r}\frac{x^r}{r!}\,, \qquad (2)$$

where $(a)_r \triangleq a(a+1)\cdots(a+r-1)$ and $(a)_0 \triangleq 1$.

For LogMMSE algorithm the estimation of $A(k,n)$ is given as

$$\hat{A}(k,n) = \sqrt{\frac{\xi_k}{\gamma_k(1+\xi_k)}}\exp\{-\frac{1}{2}(c+\sum_{r=1}^{\infty}(\frac{\gamma_k\xi_k}{1+\xi_k})^r\cdot\frac{(-1)^r}{r\cdot r!})\}R(k,n) \quad (3)$$

where c=0.57721566 is the Euler constant, the *a priori* SNR ξ_k and the *a posterior* SNR γ_k are defined as follows:

$$\xi_k \triangleq \alpha\cdot\frac{E(|A(k,n)|^2)}{E(|D(k,n)|^2)};\quad \gamma_k \triangleq \beta\cdot\frac{|R(k,n)|^2}{E(|D(k,n)|^2)}. \qquad (4)$$

In Eq.(4) α and β are the multiplicative factors which are tuned to control the balance between noise reduction and speech distortion, thus the implementation of MMSE estimation algorithms are optimized towards ASR tasks. In the experiments the optimum values for LinMMSE are α=1.00 and β=1.05, and for LogMMSE α=1.60 and β=2.13, respectively. 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. Approximation of the gain function in MMSE estimation

From Eq.(1)-(3) we can find that the gain function in LinMMSE or LogMMSE is calculated by Taylor series accumulation, which leads to a huge computational load and is a weakness of MMSE estimation algorithm especially for resource-limited embedded ASR platforms. To solve this problem, we propose to use a piece-wise linear function to substitute the Taylor series accumulation in the gain function. We take the approximation in LogMMSE algorithm as the example and describe in details as follows.

In Eq.(3) let us suppose $v \triangleq \gamma_k\xi_k/(1+\xi_k)$ and define

$$h(v) \triangleq \exp\{-\frac{1}{2}(c+\sum_{r=1}^{\infty}(\frac{\gamma_k\xi_k}{1+\xi_k})^r\cdot\frac{(-1)^r}{r\cdot r!})\}. \qquad (5)$$

Then, a suitable piece-wise linear function $pwlf(v)$ including n segments is designed to approximate the function $h(v)$ with $0 \le v \le 40$ (when v>40, $h(v) \approx 2\sqrt{\gamma_k\xi_k/(1+\xi_k)}$):

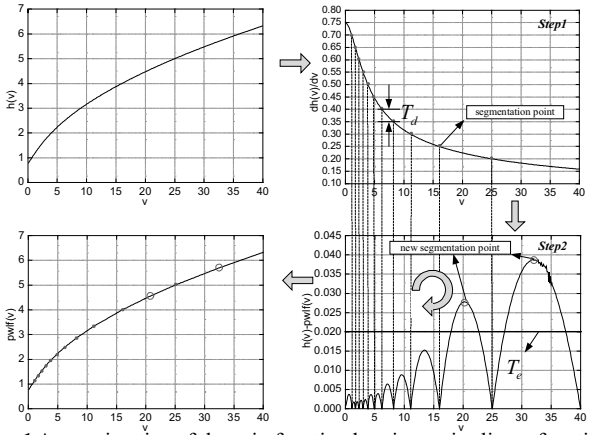

Fig. 1 Approximation of the gain function by piece-wise linear function

$$h(v) \approx pwlf(v) = \sum_{i=1}^{n} l_i(v) \quad 0 \le v \le 40\ , \qquad (6)$$

where $l_i(v)$ is the *ith* linear function between the $(i$-1$)th$ and the *ith* segmentation points of $h(v)$,denoted as $(v^{[i-1]}, h(v^{[i-1]}))$ and $(v^{[i]}, h(v^{[i]}))$ respectively:

$$l_i(v) = \begin{cases}(v-v^{[i-1]})\times\dfrac{(h(v^{[i]})-h(v^{[i-1]}))}{v^{[i]}-v^{[i-1]}}+h(v^{[i-1]}) & v^{[i-1]} \le v \le v^{[i]}\\ 0 & otherwise\end{cases}. \qquad (7)$$

Regarding the function $h(v)$ as the standard whose value is pre-calculated by Taylor series accumulation, we adopt the following steps to construct the piece-wise linear function as taking into account the derivative range of $h(v)$ and the approximation errors:

Step 1: Add initial segmentation points, which satisfy that the difference of the derivative of $h(v)$ between two consecutive points is smaller than T_d;

Step 2: if the difference between $h(v)$ and $pwlf(v)$ is greater than T_e, then insert new points in the corresponding two consecutive segmentation points;

Step 3: repeat ***Step 2*** and update $pwlf(v)$.

The above procedures are illustrated in Fig. 1 and in practice totally there are only 14 segments of linear function to approximate the gain function. It is obvious that the computation load is greatly reduced by using the proposed method.

3. SPECTRUM SMOOTHING TECHNOLOGY

The MMSE estimation algorithm can be interpreted as it suppresses or emphasizes the spectral 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.

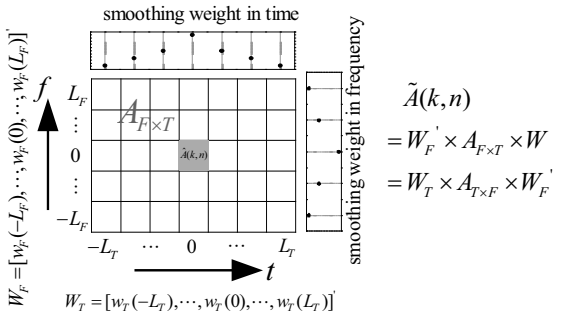

Fig. 2 Spectrum smoothing in time and frequency index

In order to compensate the spectrum components distorted by noise over-reduction, we propose to smooth the spectrum in both time and frequency index with symmetric normalized geometric 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:

$$\tilde{A}(k,n) = \sum_{i=-L_F}^{i=L_F} \sum_{j=-L_T}^{j=L_T} w_F(i) \times w_T(j) \times \hat{A}(k+i, n+j) \triangleq W_F^{'} \times A_{F\times T} \times W_T = W_T \times A_{T\times F} \times W_F^{'}, \quad (8)$$

where $w_F(i)$ is the geometric sequence weight with 0.5 common ratio in the frequency index with smoothing length $2\times L_F + 1$:

$$w_F(i) = w_F(-i) = (1 - w_F(0)) \frac{2^{L_F - i - 1}}{2^{L_F} - 1}, 1 \le i \le L_F, \quad (9)$$

$W_F = [w_F(-L_F), \cdots, w_F(0), \cdots, w_F(L_F)]^{'}$ 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.2, in Eq.(8) 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. The common ratio is equal to 0.5, which can greatly reduce the complexity in digital computation.

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. After the discrete cosine transform (DCT) decorrelation, the final 33-dimensional feature vector consists of 11 Mel frequency cepstral coefficients (MFCC) and their first and second order derivatives. To compare wit ETSI advanced front-end (ETSI_AFE) [5], we also develop a platform for evaluations on the 8KHz sampling-rate speech data.

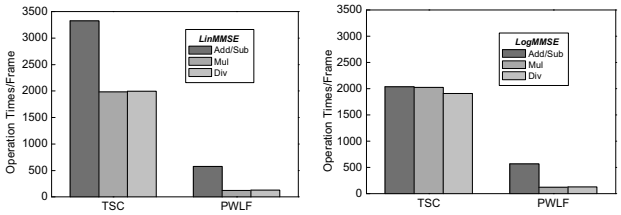

Fig. 3 Comparison of computational cost

4.2. Speech database and Acoustic model

Shanghai accented Mandarin database [14] is used to establish the isolated phrase recognition experiments to evaluate our proposed methods. We use 20000 utterances for training and 200 for evaluation. 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.

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.

4.3. Real in-car evaluation speech data

To evaluate the proposed front-end in realistic scenarios, in-car Mandarin speech data are collected from native speakers in Shanghai city. 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

5.1. Evaluations on artificial car noisy speech

Twelve car noises described in section 4.2 are used to generate the artificial evaluation noisy speech with the SNR from –5dB to 20dB. In Fig. 3 we estimate the computational cost in MMSE gain function calculation by counting the floating-point operations of addition/subtraction (Add/Sub), multiplication (Mul) and division (Div). It is obvious that the proposed piece-wise linear function approximation method (PWLF) significantly reduces the computational cost compared with Taylor series accumulation method (TSC), e.g. in LogMMSE estimation algorithm it saves about 72%, 94% and 93% computational cost in Add/Sub, Mul and Div operation, respectively.

In Fig. 4 we compare the recognition performance of different front-end schemes. We take the standard MFCC as the baseline for reference. Fig. 4(a) shows the WER averaged by 12 car noises at each SNR. We can observe that the baseline performance drops drastically when SNR is below 10dB. Applying the MMSE estimation algorithm significantly improves the robustness when compared with the baseline and it is very obvious that the spectrum smoothing algorithm further improves the recognition performance. For simplicity we only illustrate the improvement when spectrum smoothing is applied to LogMMSE algorithm, which gives the best performance in the experiments. The LogMMSE_Smooth scheme obtains the average ERR of 73.2% versus the baseline.

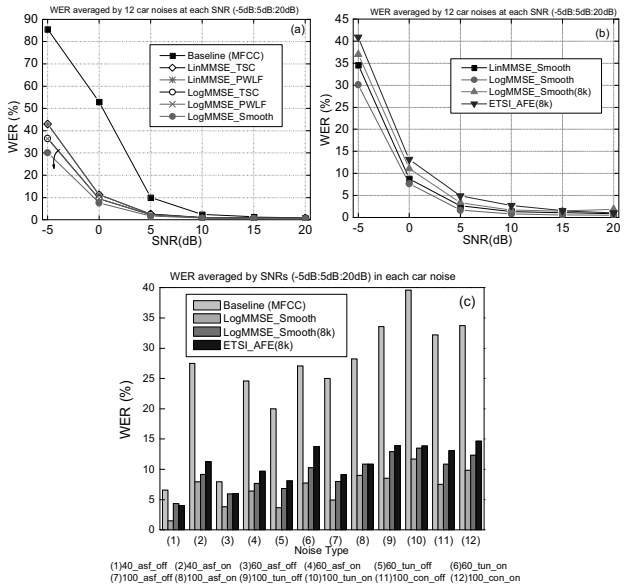

Fig. 4 Evaluation results in artificial car noisy environments

Fig. 4(a) also gives the performance of MMSE algorithm with TSC and PWLF calculation methods, respectively. We can find that using the proposed PWLF method can keep the identical recognition performance as TSC method while the computational cost is extremely reduced.

Because of the logarithmic operation in cepstral feature extraction, applying the MMSE criterion in log-spectra is better than in linear-spectra. Thus, the LogMMSE algorithm performs better than the LinMMSE algorithm as illustrated in Fig. 4(a) and Fig .4(b). When compared with ETSI_AFE, the LogMMSE_Smooth scheme obtains the average ERR of 12.2%.

Fig. 4(c) 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 speed 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 on recognition. The experimental results also show that the proposed front-end can significantly improve the performance in all conditions.

5.2. Evaluations on real in-car speech

The proposed front-end is also evaluated on real in-car speech database, as showed in Fig. 5. There are 1549 utterances in idling state test set and 1560 in driving state test set, respectively. From the experimental results, it can be concluded that the proposed front-end efficiently improves the robustness for real in-car speech recognition task and achieves the average ERR of 20.4% and 4.8% versus the baseline and ETSI_AFE, respectively.

6. CONCLUSIONS

This paper presents a robust front-end for embedded in-car speech recognition. The MMSE estimation algorithm is utilized to suppress the background noise and the gain function is approximated by a piece-wise linear function to simplify the computation complexity. A spectrum smoothing algorithm is proposed to further compensate the noise over-reduced spectra components after speech enhancement. It can be concluded from the evaluation results that the proposed front-end can efficiently improve the robustness against car noise while with low computational cost.

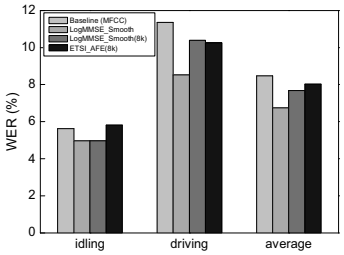

Fig. 5 Evaluation results in real in-car speech recognition

7. REFERENCES

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

[2] B. Mak, Y. Tam and Q. Li, "Discriminative auditory features for robust speech recognition", *in Proc. of ICASSP*, 2002, pp. 381-384.

[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] 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.

[5] 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.

[6] J. Droppo, L. Deng and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database", *in Proc. of Eurospeech*, 2001, pp. 217-220.

[7] W. Li, K. Itou and etc, "Adaptive regression based framework for in-car speech recognition", *in Proc. of ICASSP*, 2006, pp501-504.

[8] 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.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", *IEEE Trans. Acoustic, Speech, and Signal Processing*, Vol. ASSP-33, pp.443-445, 1985.

[10] 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.

[11] 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.

[12] 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.

[13] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms", *in Proc. of ICASSP*, 2006, pp. 153-156.

[14] X. Yan, L. He, P. Ding, R. Zhao and J. Hao, "Multi-accented Mandarin database construction and benchmark evaluations", to appear in *ISCSLP2006*.

[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.