DOI: 10.3979/j.issn.1673-825X.2012.02.001

Statistical thresholding for robust ASR

LI Yin-guo[1], PU Fu-an[1,2], Thomas Fang ZHENG[2]

(1. Automotive Electronics and Embedded Systems Engineering R&D Center,
Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China
2. Center for Speech and Language Technologies, Division of Technical Innovation and Development,
Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, P. R China)

Abstract: Speech recognition systems have been applied in real world applications for several decades, where there should be an unsatisfactory recognition performance under various noise conditions, particularly in lower signal-to-noise ratio (SNR) circumstances. In this paper, we propose a statistical thresholding method for mean and variance normalization technique, further reducing the mismatch between training and testing environments, which makes an automatic speech recognition system more robust to environmental changes. Mel-frequency cepstrum coefficient (MFCC) features are extracted as acoustic features, and they are further normalized with the mean and variance normalization method to get the cepstral mean and variance normalization (CMVN) features. The proposed statistical thresholding method is then applied. The viability of the proposed approach was verified in various experiments with different types of background noises at different SNR levels. In an isolated word recognition task, the experimental results show that the proposed approach reduced the error rate by over 40% in some cases compared with the baseline MFCC front-end, and under lower SNR conditions the proposed method also outperforms other robust features such as cepstral mean subtraction (CMS) and CMVN.

Key words: robust; feature extraction; mean subtraction; mean and variant normalization; Mel-frequency cepstrum coefficient (MFCC); statistical thresholding; speech recognition

CLC number: TN912.3　　**Document code:** A　　**Article ID:** 1673-825X(2012)02-0127-06

基于统计阈值的鲁棒性语音识别

李银国[1],蒲甫安[1,2],郑　方[2]

(1. 重庆邮电大学 汽车电子与嵌入式研究中心,重庆 400065; 2. 清华大学　语音与语言研究中心,北京 100084)

摘　要: 近几十年来,语音识别系统已由实验室环境走向真实的世界中。在不同的环境噪声下,识别性能却仍不尽人意,尤其是在低信噪比的环境中。为解决在低信噪比情况下的低识别率的问题,以声学参数 MFCC(Mel-frequency cepstrum coefficient)为基础,提出了一种基于统计阈值的倒谱均值方差归一化算法,该算法能进一步减小训练环境和测试环境的不匹配程度,从而提升了语音识别系统对环境噪声的鲁棒性。首先,对输入的语音提取 MFCC 声学参数,然后对提取的声学参数作均值方差归一化处理,最后采用统计阈值的方法抑制归一化后存在变异的特征。该算法能增加带噪语音特征和纯净语音特征的相似性;与 MFCC 为基线的系统相比,在低信噪比情况下,该算法的错误率最高下降约 40%,同时该方法也优于其他的鲁棒性特征倒谱均值减和倒谱均值归一。

关键词: 鲁棒性;特征提取;均值减;均值方差归一(MVN);梅尔频率倒谱系数(MFCC);统计阈值;语音识别

Received date: 2012-04-13
Foundation Item: The National High Technology Research and Development Program of China Project (2009ZX01038-002-002-2)

1 Introduction

A main issue in practical speech recognition is to improve the robustness against the mismatch between the train and testing environments [1]. The performance of speech recognition systems rapidly degrades if there exists channel distortion, background noise, acoustic echo, or a variety of interfering signals. A great deal of attention has been paid continuously to this problem [2], in an effort to deploy the technology in the field. When these mismatches occur, the speech recognizer could become unstable. In this paper, we will focus on the environment in which the clean speech is corrupted with background noise.

Many techniques have been deployed to alleviate the recognition performance degradation, such as model adaptation, speech enhancement, and feature compensation [3-4]. The model adaptation is powerful because there are more parameters in the system that can be changed to reduce the acoustic mismatch. Since there exists a great constraint on how these parameters can be modified, which is limited by a simple model of the degradation, this adaptation should not require a lot of speech data. Nevertheless, it has been found that due to inaccuracies in the model, the use of more data typically results in higher accuracy. An advantage of this approach is that it provides a graceful degradation under very severe mismatch conditions. One of the problems of these approaches is the large number of computations required to adapt model, which is not suitable for the rapid adaptation needed in rapidly changing environments such as telephone applications.

The speech enhancement is the simplest way to move the noise from the input noise speech signal in front-end module before feature extraction. Numerous techniques have been already studied [5]. Some of them are based on the well-known spectral subtraction (SS) approach that is suitable for enhancing speech embedded in stationary noise and its relatively simple implementation and computational efficiency. But these methods usually remain a different level of residual and unnatural background noise called musical noise. Another popular fundamental approach in speech enhancement is the Wiener Filter, which estimates clean speech from noisy speech in the sense of minimum mean square error(MMSE) given statistical characteristics. Experiments show that the amount of noise attenuation is general proportionate to the amount of speech degradation [6]. In other words, the more the noise is reduced, the more the speech is distorted. Both the musical noise and speech distortion degrade the accuracy of the recognition. So, many systems develop the approach that deployed enhanced speech for voice activity detection (VAD) and noise speech, also called unprocessed speech, for feature extraction following decoding.

Approaches that operate by compensating the input features have the limitation of having to make transformations without the acoustic knowledge used in the search process, possibly using an inaccurate correction vector. However, they typically require little computation, achieve rapid environment adaptation, and if the mismatch is not very severe they can perform as well as the model adaptation approaches do. Quite several well-known feature normalization methods for feature domain have been developed. CMS is an easy but effective way to remove the convolutional noise introduced by the transmission channel. A natural extension of CMS is cepstral mean and variance normalization (CMVN) [7]. So it can improve the robustness to additive noise as well as the channel effects. Not only does it provide with an error rate reduction under mismatch conditions, but also it has been shown to yield a small decrease in error rate under matched conditions. Those benefits, together with the fact that it is very simple to implement, is the reason why many current systems have adopted it.

In this paper, we will focus on a group of techniques called statistical matching techniques. The motivation of our statistical thresholding on mean and variance normalization (STMVN) approach is based on the following points. Firstly, as report in [8-9], modeling of speech feature distributors is discussed, which shows that speech feature distributors of each dimension can be well approximated by employing a Gaussian density model in noise environment. Secondly, STMVN has

directly physical meanings and can reduce the distance of features between the clean speech and the noise speech, which can be easily achievable

2 Statistical thresholding

In a discrete signal, if an isolated sample much different from its neighbors can be considered as impulses corresponding to high spatial frequencies. One general way to get rid of this kind of noise is to use the statistical thresholding of a small local region in the discrete signal so that out-of-range noise can be suppressed. Following is a statistical thresholding approach:

$$s[n] = \begin{cases} s[n], & \text{if } |s[n] - \mu[n]| \leqslant \sigma[n] \cdot T \\ \mu[n], & \text{else} \end{cases} \tag{1}$$

where $s[n]$ is a discrete sampled signal at moment m, T is a user specified threshold value, $\mu[n]$ and $\sigma[n]$ are the mean and the standard deviation of a small local region at m, respectively, which in turn are given by

$$\mu[m] = \frac{1}{2L+1}\sum_{n=m-L}^{m+L} s(n) \tag{2}$$

$$\sigma^2[m] = \frac{1}{2L+1}\sum_{n=m-L}^{m+L} (s[n] - \mu[n])^2 \tag{3}$$

where $2L+1$ is the length of the local region. We can see that this thresholding operation in spatial domain corresponds to low-pass filtering in the spatial frequency domain.

This approach in Eq. (1) can suppress isolated out-of-range noise, but only using the mean to substitute the out-of-range noise cannot represent the true distribution of the signal. So, we should combine Eq. (1) with standard deviation:

$$s[n] = \begin{cases} s[n], & \text{if } |s[n] - \mu[n]| \leqslant \sigma[n] \cdot T \\ \mu[n] + \text{sign}(s[n] - \mu[n]) \cdot \sigma[n] \cdot T, & \text{else} \end{cases} \tag{4}$$

where $\text{sign}(x)$ is the sign function given by

$$\text{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases} \tag{5}$$

Equation (4) shows that the isolated out-of-range noise will be substituted by mean and a rough associated with the raw data and standard deviation.

An example is shown in Fig. 1. The solid line is C4 of noise speech MFCC(mel frequency cepstrum coeficient) , and the circles are the out-of-range noise, which will be replaced by the upper or lower border (dotted line) .

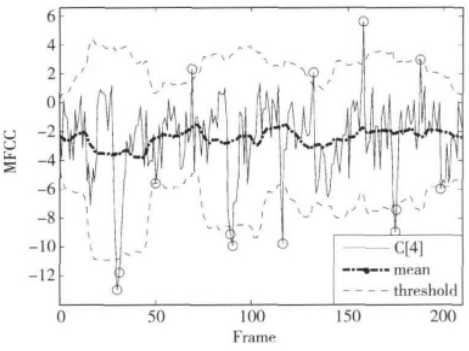

Fig. 1　C4 of noise speech MFCC

Normalizing by removing the mean shift and dividing the standard deviation on Eq. (4) , we can obtain

$$\hat{s}[n] = \begin{cases} \dfrac{s[n] - \mu[n]}{\sigma[n]}, & \text{if } \dfrac{|s[n] - \mu[n]|}{\sigma[n]} \leqslant T \\ \text{sign}(s[n] - \mu[n]) \cdot T, & \text{else} \end{cases} \tag{6}$$

where $\hat{s}[n]$ is the normalized signal.

3 Statistical thresholding on MFCC features normalization

3.1 Thresholding on MFCC features normalization

Given an ordered sequence of K-dimensional MFCC feature vectors $\boldsymbol{x}(m)$, $m = 1, \cdots, M$, then the k-th time trajectory of $\boldsymbol{x}(m)$ is denoted as

$$y_k(m) = \boldsymbol{x}(m, k), \quad m = 1, \cdots, M \tag{7}$$

where $\boldsymbol{x}(m, k)$ is the k-th component of $\boldsymbol{x}(m)$ at time m. Now, we threshold the k-th time trajectory signal $y_k(m)$ by Eq. (6) described as follow

$$\hat{y}_k[n] = \begin{cases} \dfrac{y_k[n] - \mu_k[n]}{\sigma_k[n]}, & \text{if } \dfrac{|y_k[n] - \mu_k[n]|}{\sigma_k[n]} \leqslant T \\ \text{sign}(y_k[n] - \mu_k[n]) \cdot T, & \text{else} \end{cases} \tag{8}$$

where $\mu_k(m)$ and $\sigma_k(m)$ can be obtained by Eq. (2) and Eq. (3) , respectively. We can find that Eq. (8) is similar with the traditional MVN, except the statisti-

cal thresholding operation, and it is called statistical thresholding on cepstral mean and variance normalization (STCMVN). Based on the analysis of Eq. (8), the STCMVN is split to the following two steps:

Step1 Cepstral parameters are processed by CMVN.

Step2 Thresholding operation is performed by a threshold value T.

Fig. 2 shows the C4 of clean (dashed) and noise (solid) speech CMVN. The out-of-range noise (circles) will be substituted by threshold value T or $-T$($T=2$). Fig. 3 shows a comparison of the average Euclidean distance (per frame) of CMVN and STCMVN between clean speech and noise speech. Here 1 000 voiced speeches are used to perform CMVN and STCMVN with the background of white noise at different SNR levels. We can find that the average distance of STCMVN is smaller than CMVN's at all SNR levels in Fig. 3. Combing with Fig. 2 and Fig. 3, we can find that the features of noise speech become more similar with its clean speech's after thresholding operation, alleviating the mismatch between train and testing environment.

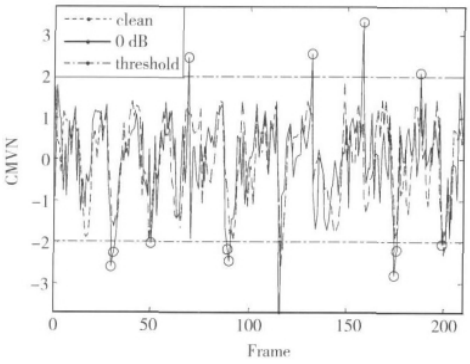

Fig. 2 C4 of CMVN

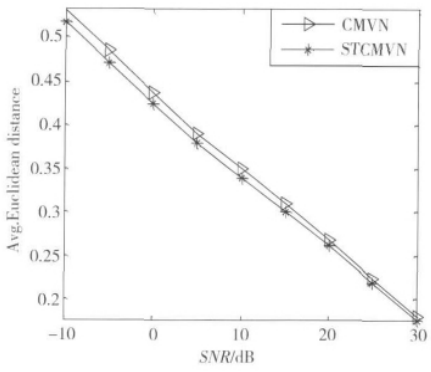

Fig. 3 Comparison of CMVN, STCMVN

3.2 Threshold determination

Threshold determination plays a vital role in our proposed approach, neither too high nor too low. We will account this situation from the statistical point of view.

In probability theory, Chebyshev's inequality, which has great utility because it can be applied to completely arbitrary distributions (unknown except for mean and variance), guarantees that in any probability distribution, no more than $1/T^2$ of the distribution's values can be more than T standard deviations away from the mean [11]:

$$P(|x-\mu| \geqslant \sigma T) \leqslant 1/T^2 . \quad (9)$$

If T is too high, there only little distribution's values more than σT, so the STCMVN will lose its effect, and if T is too small, a great number of distribution's values are out-of-range led to this value substitute by T with serious distortion. Experiments show that the choice of T is usually between 2 to 4.

4 Experimental setup and discussion

4.1 Speech corpora and setup

The speech corpora for the initial experiments included 2 600 Chinese isolate utterances produced by 31 female and 25 male speakers. The speech signal was recorded in normal laboratory environment at 16 kHz sampling rate and encoded with 16-bit quantization. In the testing phase, different background noise types, say the babble noise, the leopard noise, the pink noise, the volvo noise and the white noise, from Noisex [12] are subsequently added to the clean speech waveforms at various SNRs($SNR = -5, 0, 5, 10, 15, 20$ dB). Training utterances are not used in testing.

In speaker-independent isolated word recognition experiments, for each word with 4 states and a mixture of 6 Gaussian pdfs per state was estimated from training utterances spoken in a noise-free environment. An automatic end-pointing algorithm based on frame powers and zero crossings are used to determine the starting and ending points of the training utterances. Models are trained by CDCPM [13] which is one of the simpli-

fied HMM with the state transition left-to-right, no initial state distribution π, and state transition probability distribution A. The acoustic model employs the 26-dimension features (containing 13 MFCC coefficients plus the logarithmic frame energy, as well as their first order derivatives).

4.2 Experimental results and discussion

Experimental results for clean speech and four kinds of artificial noisy data are shown in Tab. 1 and Tab. 2-5, respectively. The clean and noisy data are evaluated by using the MFCC feature as the baseline firstly. Then the CMS, CMVN, and the proposed method ST-CMVN, with threshold value is 3.2, are applied respectively and compared.

Tab. 1　Word accuracy (%) for clean speech

Baseline	CMS	CMVN	ST-CMVN
97.24	98.48	98.29	98.38

Tab. 2　Word accuracy (%) for the Car noisy data

SNR	Baseline	CMS	CMVN	ST-CMVN
20	97.33	98.57	98.38	98.29
15	97.24	98.57	98.48	98.48
10	97.05	98.67	98.67	98.48
5	97.14	98.29	98.67	98.48
0	96.19	98.19	98.57	98.29
-5	93.43	96.57	98.29	**98.38**

Tab. 3　Word accuracy (%) for the Babble noisy data

SNR	Baseline	CMS	CMVN	ST-CMVN
20	96.95	97.71	98.10	**97.81**
15	94.57	95.33	97.43	**97.52**
10	86.19	89.52	95.43	**95.62**
5	71.14	70.95	89.14	**89.33**
0	52.29	43.62	70.19	**72.76**
-5	30.10	18.00	38.00	**42.76**

Tab. 4　Word accuracy (%) for the White noisy data

SNR	Baseline	CMS	CMVN	ST-CMVN
20	91.05	94.57	97.81	97.81
15	78.38	86.48	97.62	**97.71**
10	60.76	74.29	97.14	97.14
5	40.86	54.95	94.38	**94.48**
0	34.10	32.67	82.57	**84.00**
-5	20.10	18.76	59.43	**63.24**

Tab. 5　Word accuracy (%) for the Pink noisy data

SNR	Baseline	CMS	CMVN	ST-CMVN
20	95.05	97.43	98.10	**98.19**
15	89.81	93.62	97.71	**97.81**
10	79.90	85.62	97.14	97.14
5	56.57	62.86	95.14	94.76
0	35.71	33.52	86.57	**87.90**
-5	23.62	12.67	65.33	**68.19**

It can be seen that the proposed and CMVN approaches significantly outperform the baseline (MFCC) in all kinds of experimental environments, and the CMS approach only improves the performance of the word accuracy in higher SNR ($SNR \geqslant 10$ dB). For higher SNR noisy environments ($SNR \geqslant 10$ dB), the improvement of the proposed method is slighter with the reason is that almost all features are slightly-corrugated and the emphasized. But for lower SNR noisy environments ($SNR \leqslant 5$ dB), we can find that the proposed method has good performance than other methods except for Car noisy data because the Car noisy are stationary and narrow-band.

In particular, the STCMVN method is proved effective in the lower SNR noisy environments ($SNR \leqslant 5$ dB) where the average word accuracy for all kinds of noisy environments is more than 25%.

5 Conclusions

In this paper, an effort has been made to develop an new approach for the robust speech recognition in noisy environments by using a statistical threshold value to threshold the CMVN features, reducing the mismatch between train and testing environments. The experimental results show that the proposed approach is superior over the other robust features (CMS, CMVN), especially for lower SNR cases (-5 dB $\leqslant SNR \leqslant$ 5 dB).

References:

[1] GONG Yifan. Speech Recognition in Noise Environments: A Survey [J]. Speech communication, 1955, 16(3): 261-291.

[2] ACERO A. Acoustical and Environmental Robustness in Automatic Speech Recognition [M]. UK: Kluwer Aca-

demic Publishers, 1993.

[3] HUANG X D, ACERO A, HON X. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development [M]. New Jersey, USA: Prentice Hall, 2001.

[4] VIIKKI Olli, LAURILA K. Cepstral domain segmental feature vector normalization for noise robust speech recognition [J]. Speech Communication, 1998, 25(1-3): 133-147.

[5] ACERO A. Acoustical and Environmental Robustness in Automatic Speech Recognition [D]. Pittsburgh: Carnegie Mellon University, 1990.

[6] BENESTY J, MAKINO S, CHEN J. Speech Enhancement, Signal And Communication Technology [M]. UK: Springer-Verlag Berlin and Heidelberg GmbH & Co. K, 2005.

[7] CHEN Chia-Ping, BILMES J A. MVA Processing of Speech Features, IEEE Transactions on Audio [J]. Speech and Language Processing, 2007, 15(1): 257-272.

[8] GAZOR S, ZHANG W. Speech Probability Distribution [J]. IEEE Signal Processing Letters, 2003, 10(7): 204-207.

[9] DU Jun, WANG R H. Cepstral Shape Normalization (CSN) For Robust Speech Recognition [C] // Proc. of ICASSP, [s. l.]: IEEE Press, 2008: 4389-4392.

[10] Internet Center for Management and Business Administration, Inc. Statistics [EB/OL]. [2012-03-10]. http://www.quickmba.com/stats/dispersion.

[11] Chebyshev's inequality, Wikipedia [EB/OL]. [2012-03-11]. http://en.wikipedia.org/wiki/Chebyshev's_inequality

[12] VARGA A P, STEENEKEN H J M, Tomlinson M, et al. The NOISEX-92 study on the effect of additive noise on automatic speech recognition [R]. Malvern, UK: Speech Research Unit, Defense Research Agency, 1992.

[13] ZHENG Fang, CHAI H-X, SHI Z-J. A Real-World Speech Recognition System Based on CDCPMs [C] // 971nt. Conf. Computer Processing of Oriental Langunges (ICCPOL'97), Apr. 2, 1997, Hong Kong: [s. n.], 1997: 204-207.

Biographies:

Li Yin-Guo (1955-), Professor, Hupei, His main research directions are the pattern recognition and artificial intelligence, System identification and intelligent control. E-mail: liyg@cqupt.edu.cn.

Pu Fu-An (1986-), graduate, Sichuan, His main research directions are the pattern recognition and robust Automatic Speech recognition. E-mail: pufa@cslt.riit.tsinghua.edu.cn.

Thomas Fang ZHENG (1967-), Professor, Jiangsu, his research interest is speech and language processing. He is an IEEE Senior Member, a BoG member of APSIPA, a council member of Chinese Information Processing Society of China, a council member of the Acoustical Society of China. He is an associate editor of IEEE Transactions on Audio, Speech, and Language Processing, a member of editorial board of Speech Communication, a member of editorial board of APSIPA Transactions on Signal and Information Processing, an associate editor of International Journal of Asian Language Processing. E-mail: fzheng@tsinghua.edu.cn.

(编辑:魏琴芳)