频域动态特征提取中的多层信道正规化*

王 东+, 朱小燕, 刘 盈

(清华大学 计算机科学与技术系,北京 100084)
(清华大学 智能技术与系统国家重点实验室,北京 100084)

Multi-Layer Channel Normalization for Frequency-Dynamic Feature Extraction

WANG Dong+, ZHU Xiao-Yan, LIU Ying

(Department of Computer Science and Techology, Tsinghua University, Beijing 100084, China)
(State Key Laboratory of Intelligent Technology and System, Beijing 100084, China)
+ Corresponding author: E-mail: wangdong99@mails.tsinghua.edu.cn
http://www.tsinghua.edu.cn
Received 2002-06-24; Accepted 2002-09-23

Wang D, Zhu XY, Liu Y. Multi-Layer channel normalization for frequency-dynamic feature extraction. *Journal of Software*, 2003,14(9):1523~1529.
http://www.jos.org.cn/1000-9825/14/1523.htm

Abstract: Despite the steady progress made in the area of speech recognition and a high number of practical applications, it is widely acknowledged that recognition technology today is not at the desired level. One main obstacle is what said "robustness". This paper focus on one popular idea in antagonizing speech system vulnerability-channel normalization, and presents a new normalization algorithm MLCN (multi-layer channel normalization), which exploits the recursive compensation progress in two domains (spectral domain and cepstral domain) to depress the noise and channel distortion, so that the more robust speech representation for the following processing is achieved. A new frequency-dynamic feature extraction algorithm is also proposed due to the introduction of MLCN, which allows dynamic information integrated in the final feature vectors. Experimental results of the gallina system demonstrate the validity of the new algorithm.
Key words: speech recognition; feature extraction; MFCC; channel normalization; frequency-dynamic feature

摘 要: 语音识别领域已经取得了稳步发展并出现了众多实用系统,但众所周知,今天的识别技术还远没有达到要求,而"鲁棒性"问题是系统性能提高的一个主要障碍.集中讨论了一种对抗语音识别系统脆弱性的通行方法——信道正规化技术,提出了一种新的正规化策略——多层信道正规化 MLCN(multi-layer channel

* Supported by the National Natural Science Foundation of China under Grant No.69982005 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.G199803050703 (国家重点基础研究发展规划项目(973))

WANG Dong was born in 1975. He is a master student at the State Key Laboratory of Intelligent Technology and System, Tsinghua University. His research interest is spoken language processing. **ZHU Xiao-Yan** is a professor at the Department of Computer Science and Technology, Tsinghua University. His research interests are artificial intelligence and signal processing. **LIU Ying** was born in 1980. She is a master student at the Department of Computer Science and Technology at the Tsinghua University. Her interest is very large vocabulary speech recognition.

normalization)新的算法应用递归补偿算法,在频谱域和倒谱域两层上进行正规化,降低噪音和去除信道畸变,从而为后续识别过程提供更鲁棒的特征参数.在此基础上,探讨了一种新的语音识别特征参数的提取-频域动态倒谱系数,由于 MLCN 的引入,频域的动态信息被恰当地集成到最终的特征向量中.在 gallina 系统中的实验证明了这种新方法的有效性.

关键词: 语音识别;特征提取;Mel 倒谱系数;信道正规化;频域动态特征

中图法分类号: TP391 **文献标识码**: A

Speech recognition technology is regarded as one of the prominent subjects in the coming five years, due to its convenience, naturalness, and humanization. Ironically, although some practical systems have served successfully in various applications, the real-world performance is far from the people's desired level. One main widely acknowledged obstacle is the vulnerability introduced by applying conditions, including acoustic variability, noise background, domain discrepancy, speaker distinctness, and so on. Many strategies have been developed in recent years to counterwork the distortion with various perspectives, such as feature extraction, model construction and decoder set-up. In front-end, almost all the early methods work in spectral domain and fall into three classes - normalization, adaptation, and corruption-immunity, which still guide the nowadays research.

Since MFCC became the standard front-end, much effort has been taken to improve its efficiency in real-world environment, such as cepstral mean subtraction (CMS)[1], distortion-constraint strategies including multi-band feature extraction and acoustic backing-off decoder[2,3], and alternative linear transforms such as PCA and LDA[4,5], instead of normal Discrete Cosine Transform (DCT). Other types of front-ends as LSF[6], which try to incorporate some of the features of the human auditory mechanism, have been suggested for robust speech recognition. Combined with RASTA[7] processing, these front-ends achieved better performance in noisy conditions compared with MFCC.

Besides the feature extraction, some noise-compensative or noise-resistant strategies on level of acoustic models are also supposed, such as DPMC[8] and spectral addition model[9]. Some classical adaptive methods as MAP and MLLR can also be regarded as noise-compensative methods.

Our direct motivation of this research is the great discrepancy in our databases in which not only the training-test mismatch but also the within-training incompatibility is serious, so that an effective and real-time normalization method is desired. Spectral subtraction and cepstral mean normalization are considered because of their simplicity, but the result of either algorithm is unsatisfactory, which inspires us to combine these two methods together.

Another momentum comes from our desire to integrate the spectral-dynamic information into the feature vector, just as time-dynamic components in MFCC. Although this idea has been presented in some literatures[10], hereto few successful instances have been reported. Analysis of our experimental results reveals that the unbalance among various information sources is necessary and important. It's fortunate that the normalization in spectral domain just provides such balance, which also encourages us to integrate the normalizations in both spectral and cepstral domains together.

Thanks for the recursive framework introduced by Olli Viikki[11], we can apply the similar strategy in spectral domain too, which extends "NSS" (Noise Spectral Subtraction) to "SMN" (Spectral Mean Normalization) so that the real-time compensation is available, without using silence detector.

On the considerations above, a new Multi-Layer Channel Normalization (MLCN) is presented in recursive framework. The new algorithm combines the normalizations in spectral and cepstral domains, achieving more significant error reduction.

In the next section, the new algorithm will be illustrated in detail, and then the frequency-dynamic feature

extraction is presented in Section 2, where the validity to combine various information sources, in our case the static and dynamic frequency banks, is explained based on the new normalization. Our experiments are given in Section 3, and the main idea will be concluded in the last section.

1 Multi-Layer Channel Normalization

1.1 Spectral and cepstral normalization

As discussed above, conventional channel normalization methods are implemented in spectral or cepstral domain separately, and since the purpose is consistent, only one is chosen, as illustrated as [a] and [b] in Fig.1. In Fig.1, the diagram Pre-Processing contains high-frequence emphasizing, frame splitting, FFT and triangle energy filtering on Mel axis. The outputted Mel banks are in spectral domain, and will be translated to cepstral domain through a linear transform, such as Distributed Cosine Transform (DCT), so that MFCC is generated. It should be noticed that, unlike conventional algorithms as [a] and [b], in our new strategy, normalizations in two domains are both important.

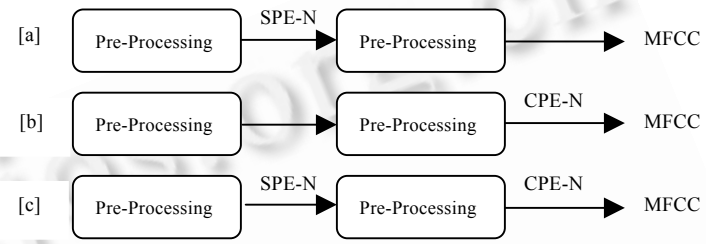

Fig.1 Normalizations in spectral and cepstral domains

1.2 Multi-Layer channel normalization

It's well known that the distortion of speech signal mainly comes from two aspects, noise and channel affection. Although additive and convoluted noise all exist in real condition, the later can be regarded as a noisy channel in mathematics since its stationary and property of convolution in temporal domain. Then the real-word speech $x(k)$ could be formulated as following

$$x(k)=[v(k)+n(k)]\otimes h(k), \tag{1}$$

where $\otimes$ is the operator of convolution, $v(k)$ and $n(k)$, the clean speech and additive noise respectively, and $h(k)$ is the convoluted noise and channel affection.

Translated to spectral domain, the additive noise will be separated from speech, as follows

$$spe_x(m)=spe_v(m)spe_h(m)+spe_n(m)spe_h(m), \tag{2}$$

where $spe(m)$ denotes the corresponding spectrum of each signal component. Since noise and channel are both stationary, the long-term average of the spectrum will leave only the second item of the right side of equation (2). Conventional noise spectral subtraction (NSS) deletes this additive noise and gets the cleaned speech as

$$spe'_x(m)=spe_v(m)spe_h(m). \tag{3}$$

In our experiments, a modified recursive strategy suggested in Ref.[11] is applied in spectral domain, which not only avoids the need to design the voice active detector, but also updates the compensation factors in real time. Equation (4) gives the normalization procedure.

$$\hat{spe}_x(m)=\frac{spe_x(m)-u_m}{\sigma_m}, \tag{4}$$

where u_m is the long-term average of spectrum $spe_x(m)$ and σ_m is the covariance. According to equation (3), we get

$$\hat{spe}_x(m)=\sigma_m^{-1}spe_v(m)spe_h(m). \tag{5}$$

It shows that additive noise has been eliminated from original spectrum. What should be noticed is that, the magnitude of each $\hat{spe}_x(m)$, m=1,2,…,M is also normalized by the covariance, which is important in the next section.

To calculate the compensation factors u_m and σ_m, recursive procedure is applied. Here two steps are important: initialization and update of compensation factors. In our experiments, first T frames are used to estimate the initial compensation factors, applying equations (6) and (7),

$$u_m = \frac{1}{T}\sum_{t=1}^{T} spe_x(m;t)\,, \tag{6}$$

$$\sigma_m = \sqrt{S_m} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}[spe_x(m;t)]^2 - u_m{}^2}\,, \tag{7}$$

where $spe_x(m;t)$ is the *m-th* spectral component of the *t-th* frame of signal x. Then no matter whether the next coming frame is in the initializing set or not, it will be normalized by the current factors using equation (4), and will be used to update the factors as follows,

$$\hat{u}_m = au_m + (1-a)spe_x(m;t)\,, \tag{8}$$

$$\hat{S}_m = aS_m + (1-a)[spe_x(m;t)]^2\,, \tag{9}$$

where a is an adaptation-factor whose value is determined experimentally according to the number of initial frames T.

Normalized in spectral domain, the spectrum is compressed and translated to cepstral domain by a linear transform, such as DCT in MFCC generation. Combined with (5), the cepstral coefficients are described by (10),

$$CEP_x = LT\{\log(\Sigma^{-1}SPE_v)\} + LT\{\log SPE_h)\}\,, \tag{10}$$

with

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \cdots & \\ & & & \sigma_M \end{bmatrix},$$

where LT is certain linear transform, SPE, the spectrum vector. CEP_x is the cepstral coefficient vector, each dimension of which can be written as

$$cep_x(q) = cep'_v(q) + cep_h(q)\,, \tag{11}$$

where $cep_h(q)$ is the *q-th* cepstral coefficient of convoluted noise and channel distortion, and $cep'_v(q)$ corresponds the cepstral coefficient of clean speech, but a factor σ^{-1} is included.

The similar recursive algorithm described in (4)~(9) is also exploited in cepstral domain in our experiments, and as many literatures have proved, this normalization cancels the distortion component $cep_h(q)$ in equation (11). Hence the final cepstral feature is

$$\hat{cep}_x(q) = \frac{ceq_x(q) - u'(q)}{\sigma'(q)}\,, \tag{12}$$

where $u'(q)$ is the average of the *q-th* cepstrum, and $\sigma'(q)$, the covariance. Obviously these two normalizations introduced by (4) and (12) contribute differently: spectral normalization deletes additive noise, and cepstral normalization eliminates the convoluted distortion, including noise and channel affection.

As discussed above, the recursive multi-layer normalization algorithm is designed as follows.

(i) Spectral normalization factors initialized.

(ii) Cepstral normalization factors initialized.

(iii) For each frame

Spectral normalization implemented using Eq.(4).

Spectral normalization factors updated.

Raw cepstral coefficients generated through linear transform, i.e., DCT.

Cepstral normalization implemented using Eq.(12).

Cepstral normalization factors updated.

Frame End.

2 Frequency-Dynamic Feature Extraction

It's well known that time-dynamic features provide significant performance increase in speech recognition , so it's natural to consider whether the frequency-dynamic information could be helpful. In fact, in MFCC generation, overlay of neighboring banks on Mel axis is an implicit integration of this dynamic information. More explicit consideration is suggested in Ref.[10], where a filter $H(z) = z - z^{-1}$ is applied upon original Mel banks to represent the dynamic information.

Referring to Ref.[10], we use $H(z)$ to generate delta filter-bank energies (delta-FBEs), and combine them with the first and last static FBEs to format the feature vector, but got much worse performance compared with MFCC. Analysis shows that, the high error rate may come from the unorthodoxy between dimensions of the feature vector. As we know, standard HMM uses Gaussian density distributions with diagonal covariance matrix to describe the state emission, which requires the orthodoxy among feature dimensions.

To ensure the orthodoxy, we use a linear transform to convert FBEs to cepstral coefficients, but different from MFCC, DCT is substituted by discrete Karhunen-Loeve transform, because the sequence of FBEs may not be approximated with a first-order Markov model. The experimental result shows that this transform indeed increases the performance, but unfortunately, the result is still unsatisfactory, and it seems that the frequency-static and dynamic information haven't been integrated properly.

Further analysis reveals that, it's the static FBEs that mask the dynamic ones. Because of the continuity in frequency domain, dynamic Mel banks through filter $H(z)$ are on much lower significant level than the static components, so the balance among Mel banks is desired to boost the contribution of dynamic FBEs.

As shown in (4), the spectral normalization not only deletes the additive noise, but also provides a balance factor σ that normalizes the magnitude of each spectral dimension, and this factor will be retained until cepstral normalization as expressed in (10)~(12). So the spectral normalization in the multi-layer algorithm is a good method to balance spectral dimensions, and with the following linear transform and cepstral normalization, the orthodoxy and noise immunity are ensured respectively. As will be seen in the following section, the new algorithm not only provides satisfactory performance improvement for frequency-dynamic feature, but also provides much more than for the normal MFCC.

3 Experimental Results

All the experiments were progressed in our gallina continuous speech recognition system. Current gallina system uses MFCC as the front end, and continuous HMM as its acoustic model, with time-synchronous viterbi decoder. Two databases are used to test the proposed algorithm: isolate syllable database CIDS and continuous speech database 863CSL. The former database is recorded in consistent environment, while the later one contains obvious discrepancy in acoustic environment since it's recorded in two steps. We will see different contributions of our proposed algorithm.

Experiment A: Contribution of multi-layer normalization to conventional MFCC.

In the following series of experiments, CIDS is firstly employed to set up 411 un-toned Chinese syllable models, in which 5 states with 8 Gaussian mixtures are concatenated, allowing one state jump-over.

Totally about 1322 toned isolate syllables from 40 males are gathered to train the isolate syllable models, and the left 20 persons' utterances test the results.

Based on the isolate syllable models, continuous training is progressed using totally 20,000 sentences of 40 male speakers in 863CSL, and then, 10 speakers are used to test the model. To examine the contribution of each type of normalization, four experiments are carried out for each database respectively, as show in Fig.2, where SMN represents the normalization in spectral domain, and CMN, in cepstral domain. In Fig.2, Word Error Rate (WER) is given for each case.

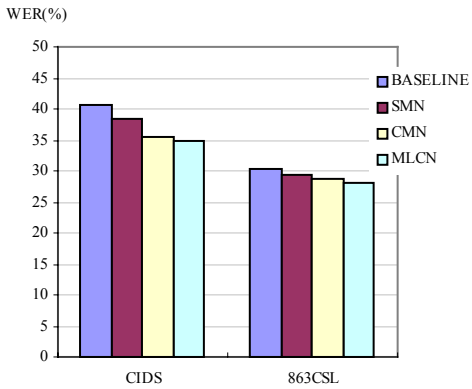

Fig.2 Contribution of MLCN to MFCC

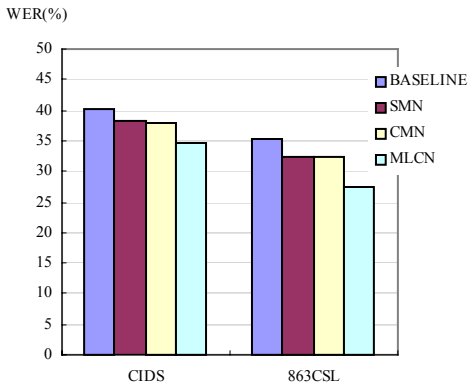

Fig.3 Contribution of MLCN to frequency-dynamic feature

It can be seen that: (1) Normalizations in each domain provides significant performance improvement, and the cepstral one seems contributes more salient. This result is expected since all Mel banks in MFCC are almost in the same magnitude level, so that the effect of spectral normalization is only additive noise depression; (2) The combination of two normalizations indeed provides more significant error reduction than any separate one, which is more obvious in continuous case because of the more channel distortion in 863CSL.

Experiment B: Simple test for frequency-dynamic features.

In this experiment we will propose a test for frequency-dynamic feature introduced by Ref.[10], where the filter $H(z)=z-z^{-1}$ is used to generate the delta-FBEs as final features. In the simple test, only 40 syllable models are trained using materials of 20 persons in CIDS, and other 20 persons test. Eliminating energy and time dynamic components, the feature vectors contain only time-static information. Six types of features, including what suggested in Ref.[10] and the ones derived from it are tested, as following list:

T1: FBEs+ $H(z)$, the first and last static FBEs are copied (P2 in Ref.[10])

T2: T1+K-L transform

T3: T2, but the first and last spectral FBEs are set to 0

T4: T2+spectral normalization.

T5: T2+cpestral normalization.

T6: T2+multi-layer normalization.

Table 1 Simple tests of various frequency-dynamic features

Feature	T1	T2	T3	T4	T5	T6
Error rate(%)	42.7	39.7	39.1	37.01	34.24	31.94

It should be noticed that T2 achieves worse performance than T3, although two static banks are added. This can be explained by the masking effect of the static banks upon the dynamic ones. It's also clear that, by the normalization of each bank, the balanced features obtain much better performance, as indicated by T4. At last, as T6 shows, the integration of the two normalizations achieves the most significant error reduction.

Experiment C: Contribution of multi-layer normalization to frequency-dynamic features.

In this experiment, T6 in experiment B is examined thoroughly in the same condition of experiment A. The result shown in Fig.3 certifies the validity of our analysis, indicating that, the multi-layer normalization contributes much than in experiment A, because the effect of normalization in spectrum is not only deleting the noise, but also balancing the Mel banks. Also, it can be seen that, because more information is combined properly, new feature achieves slight performance improvement than MFCC.

4 Conclusions

In this paper, a new multi-layer channel normalization idea is proposed and implemented in recursive adaptation framework. A series of comparative experiments demonstrate the effectuality of this new algorithm. On the other hand, a modified frequency-dynamic feature extraction algorithm is presented, and the remarkable contribution of our new algorithm to this new feature is testified and analysed. The experimental results indicate that, features extracted from multi information sources can work only when proper balance among these sources exists.

References:

[1] Liu FH, Stem RM, Acero A, Moreno PJ. Environment normalization for robust speech recognition using direct cepstral comparison. In: Proceedings of the ICASSP94. 1994. 61~64.

[2] de Veth J, Cranen B, Boves L. Acoustic backing-off as an implementation of missing feature theory. Speech Communication, 2001,34:247~265.

[3] Hariharan R, Kiss M, Viikki O. Noise robust speech parameterization using multiresolution feature extraction. IEEE Transactions on Speech and Audio Processing, 2001,9(8):856~865.

[4] Hermansky H. Spectral basis functions from discriminant analysis. In: Proceedings of the ICSLP'98. 1998. 1379~1383.

[5] Hunt M, Bateman D, Richardson S, Piau P. An investigation of PLP and IMELDA acoustic representation and of their potential for combination. In: Proceedings of the ICASSP'91. 1991. 881~884.

[6] Paliwal KK. A study of LSF representation for speaker-dependent and speaker independent HMM-based speech recognition systems. In: Proceedings of the ICASSP'90. 1990. 801~804.

[7] Hermansky H, Morgan N. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing, 1994,2(4):578~579.

[8] Gales MJF, Young SJ. A fast and flexible implementation of parallel model combination. In: Proceedings of the ICASSP'95. 1995. 133~136.

[9] Sanches I. Noise-Compensated hidden Markov models. IEEE Transactions on Speech and Audio Processing, 2000,8(5):533~540.

[10] de Veth J, de Wet F, Cranen B, Boves L. Acoustic features and a distance measure that reduce the impact of training-test mismatch in ASR. Speech Communication, 2001,34:57~74.

[11] Viikki O, Bye D, Laurila K. A recursive feature normalization approach for robust speech recognition in noise. In: Proceedings of the ICASSP'98. 1998.