A NEW NORMALIZATION FOR MFCC: MULTI LAYER STRATEGY AND RECURSIVE PROGRESS

Wang, Dong; Zhu, Xiaoyan; Liu, Ying

State Key Laboratory of Intelligent Technology and System
Department of Computer Science, Tsinghua Univ.

Wangdong99@mails.tsinghua.edu.cn
Zxy-dcs@tsinghua.edu.cn

Abstract

One main obstacle in speech recognition is what said “robustness”. This paper focus on one popular idea in antagonizing speech system vulnerability-channel normalization, and presents a new normalization algorithm- Multi-Layer Channel Normalization (MLCN), which exploits the recursive compensation progress in two domains- spectral domain and cepstral domain- to depress different noises, so that the more robust speech representation is achieved. Experimental results of our gallina system demonstrate the validity of our new algorithm.

Key Words: Speech Recognition, Feature Extraction, MFCC, Channel Normalization,

1. INTRODUCTION

Although some practical systems of speech recognition 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, successful methods of speech enhancement include: spectral subtraction introduced by Boll [1]; signal restoration by spectral mapping suggested by Juang and Rabiner [2]; adaptive filtering techniques as Kalman filter [3] and all-poll modeling of degraded speech, proposed by Lim and Oppenhein [4]. Almost all these early methods work in spectral domain and fall into three classes - normalization, adaptation, and corruption-immunity, which still guide the nowadays research.

Since MFCC was introduced by Davis and Mermelstein [5] and became the standard front-end, much effort has been taken to improve its efficiency in real-world environment, such as cepstral mean subtraction (CMS) [6][7], distortion-constraint strategies including multi-band feature extraction and acoustic backing-off decoder [8][9], and alternative linear transforms such as PCA and LDA [10][11], instead of normal Discrete Cosine Transform (DCT). Other types of front-ends as LSF [12] and PLP [13], which try to incorporate some of the features of the human auditory mechanism, have been suggested for robust speech recognition. Combined with RASTA [14] 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 PMC [15], DPMC [16] and spectral addition model [17]. Some classical adaptive methods as MAP and MLLR can also be regarded as noise-compensative methods.

With the recursive framework introduced by Olli Viikki [18], 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. In this case, a multi-layer channel normalization will be proposed and the experimental results show the validity of this method.

Next section will introduce our new algorithm in detail, which will be followed by the experimental results in section 3, and at last, the whole idea will be concluded in the last section.
2. MULTI-LAYER CHANNEL NORMALIZATION

2.1 Spectral and cepstral normalization

As discussed in section 1, 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 figure 1.

In figure 1, the diagram Pre-Processing contains high-frequency 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.

2.2 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

\[
x(k) = [v(k) + n(k)] \otimes h(k)
\]

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

\[
spe_x(m) = spe_x(m)spe_y(m) + spe_y(m)spe_x(m)
\]

Where \( spe_x(m) \) 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

\[
spe'_x(m) = spe_x(m)spe_y(m)
\]

In our experiments, a modified recursive strategy suggested in [20] 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.

\[
spe'_x(m) = \frac{spe_x(m) - u_m}{\sigma_m}
\]

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

\[
spe'_x(m) = \sigma_m^{-1} spe_x(m) spe_y(m)
\]

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

To calculate the compensation factors \( u_m \) and \( \sigma_m \), 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 equation (6)(7),

$$u_m = \frac{1}{T} \sum_{t=1}^{T} \text{spe}_x(m;t)$$  \hspace{1cm} (6)

$$\sigma_m = \sqrt{S_m} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (\text{spe}_x(m;t))^2 - u_m^2}$$  \hspace{1cm} (7)

where $\text{spe}_x(m;t)$ 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,

$$\hat{u}_m = au_m + (1-a)\text{spe}_x(m;t)$$  \hspace{1cm} (8)

$$\hat{S}_m = aS_m + (1-a)(\text{spe}_x(m;t))^2$$  \hspace{1cm} (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),

$$\text{CEP}_x = LT\{\log(\Sigma^{-1}\text{SPE}_x)\} + LT\{\log\text{SPE}_h\}$$

$$\Sigma = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_M \end{bmatrix}$$  \hspace{1cm} (10)

where $LT$ is certain linear transform, $\text{SPE}$, the spectrum vector. $\text{CEP}_x$ is the cepstral coefficient vector, each dimension of which can be written as

$$\text{cep}_x(q) = \text{cep}_x(q) + \text{cep}_h(q)$$  \hspace{1cm} (11)

where $\text{cep}_h(q)$ is the $q$-th cepstral coefficient of convoluted noise and channel distortion, and $\text{cep}_x(q)$ corresponds the cepstral coefficient of clean speech, but a factor $\sigma^{-1}$ 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 $\text{cep}_h(q)$ in equation (11). Hence the final cepstral feature is

$$\hat{\text{cep}}_x(q) = \frac{\text{cep}_x(q) - \mu(q)}{\sigma(q)}$$  \hspace{1cm} (12)

where $\mu(q)$ 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.

1. Spectral normalization factors initialized.
2. Cepstral normalization factors initialized.
3. For each frame
   - Spectral normalization implemented using equation 4.
   - Spectral normalization factors updated.
   - Raw cepstral coefficients generated through linear transform, i.e., DCT.
   - Cepstral normalization implemented using equation 12.
   - Cepstral normalization factors updated.
4. Frame End.

### 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.

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 figure 2, where SMN represents the normalization in spectral domain, and CMN, in cepstral domain. In figure 2, Word Error Rate (WER) is given for each case.

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.

4. CONCLUSION

In this paper, a new multi layer channel normalization idea is proposed and implemented in recursive adaptation framework. The experimental results suggest that, normalizations in two domains work differently and provide more error reduction.

5. REFERENCE


