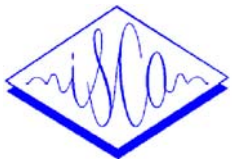

4th International Conference on Spoken Language Processing (ICSLP 96)
Philadelphia, PA, USA
October 3-6, 1996

# A Unified Spectral Transformation Adaptation Approach for Robust Speech Recognition

*Lei Yao, Dong Yu & Taiyi Huang*

email: huang@prldec3.ia.ac.cn
National Lab of Pattern Recognition, Institute of Automation,Chinese Academy of Sciences
Beijing 100080, China

## ABSTRACT

In this paper, Canonical Correlation Based Compensation(CCBC) is proposed as an unified approach to cope with the mismatch between training and test set. The mismatch between training and test conditions can be simply clustered into three classes: differences of speakers, changes of recording channel and effects of noisy environment. In previous work, we had used CCBC approach with some modifications to make our speech recognizer robust to the noisy environment successfully[1]. Recently, the same approach has been extended for speaker and channel adaptation. The results of our experiments show that CCBC approach well compensated all three kinds of distortion source between training and test conditions. In order to compare the performance of CCBC with that of some conventional adaptation approaches, the capacities of the techniques of cepstral mean normalization, RASTA and Lin-Log RASTA are tested. We find that CCBC has better performance than them. As an very important problem in CCBC approach, the selection of appropriate reference speech data is also discussed in this paper.

## 1. INTRODUCTION

Over the past decade, we have witnessed that speech recognition in controlled situations has reached very high level of performance. However, the deployment of speech recognition technology is still hampered by lack of robustness in system performance. It is common to have a recognition system's error rate increase by several folds when tested using a microphone different from the one on which it was trained. Similarly, degradation in recognition performance are often observed when the system is used by a new speaker. In case of high noisy environments, the recognition system will often be corrupted to be unacceptable.

Channel, speaker and additive noise affect the speech signal in different ways. Convolutional distortion may be introduced by speakers' vocal tracts and microphone transfer functions. The acoustic variations affected by noisy environment come from two ways. First, additive noise contaminates the speech signal and changes the characteristic vectors representing speech. Second, when the speaker attempts to increase the communication efficiency over the noisy medium, speaking causes statistically significant articulation variability. This is known as lombard effect.

Some previous researchers[3][5][6][8][12] had performed compensation for three kinds of distortion sources mentioned above. But their studies had coped with each one with a different processing approach. In this paper, we utilized an unified spectral transformation adaptation method to compensate all three kinds of distortion sources affecting the speech signal. Unlike EM algorithm commonly used in spectral transformation compensation, which is an iterative one, CCBC is proved to have a solution. Its calculating procedure is specific and short.

Compared with training speech, the cepstrum of test speech has three main changes affected by all kinds of distortion sources: mean value shift, norm shrink and the bad correlation of each dimension between training vectors and test vectors[1]. While other compensation methods often deal with only one or two of the cepstrum changes, for example, cepstral normalization only compensates the cepstral mean value shift, CCBC as a linear cepstral transforming approach amounts to a rotation and scaling in cepstral vector space. And it reconstructs the correct correlation between training vectors and test vectors. So it can compensate all three kinds of cepstral variations affected by distortion sources and can be an unified spectral transformation adaptation approach to deal with all kinds of mismatch between training and test set.

In Section 2 we describe the algorithm of CCBC. In section 3 we report the performance of CCBC, cepstral mean normalization, RASTA and Lin-Log RASTA on our speaker-independent VQ/DHMM isolated-word speech recognition system. Finally, in Section 4 we present our conclusions.

## 2. CANONICAL CORRELATION BASED COMPENSATION

### 2.1. Algorithm

Speech signal can be represented as a sequence of feature vectors, each vector can be thought as a point in the feature vector space. In our case, we used P-order mel-frequency cepstral coefficients as the feature vector. The differences between training vectors and test vectors can be compensated by CCBC. But CCBC does not directly transform test space to training space. It makes the training vectors and test vectors maximum correlation in the reference space (the third space). If we regard that training vectors and test vectors are the vectors $X^{(1)}$ and $X^{(2)}$ respectively, we can suppose:

$$U=AX^{(1)}, V=BX^{(2)}$$

where A and B are the transformation matrixes corresponding to $X^{(1)}$ and $X^{(2)}$ respectively, U and V are the mappings of $X^{(1)}$

and $X^{(2)}$ in the reference space. We minimize the mean-square-error

$$D= E\{(U-V)^2\}$$

with constraints $E\{U^2\}=E\{V^2\}=1$. That is we make U and V maximum correlation and assure that U and V can not be zero at the same time. We solve this problem by the following procedure.

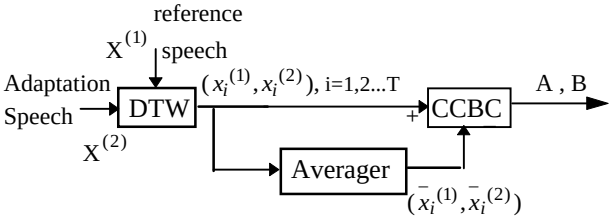


Figure 1: Procedure of Calculating the transformation of CCBC

We assume that $X=\begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}$. With the supposition that the long-time mean of speech cepstrum is zero, we can make $E\{X\}=0$ if we subtract the channel characteristics from the training and test vectors respectively. That is we can get $E\{X^{(1)}\}=E\{X^{(2)}\}=0$ if $X^{(1)}= X^{(1)}-\overline{X}^{(1)}$ and $X^{(2)}= X^{(2)}- \overline{X}^{(2)}$, Thus we get the correlation matrix : $\Sigma=\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ and the relations:

$$1= E\{U^2\}=E\{A' X^{(1)} X^{(1)'} A\}= A' \Sigma_{11} A$$

$$1=E\{V^2\}=E\{B' X^{(2)} X^{(2)'} B\}= B' \Sigma_{22} B$$

$$E\{U\}=E\{A' X^{(1)}\}=A' E\{X^{(1)}\}=0$$

$$E\{V\}=E\{B' X^{(2)}\}= B' E\{X^{(2)}\}=0$$

$$E\{UV\}=E\{A' X^{(1)} X^{(2)} B\}= A' \Sigma_{12} B$$

The problem can be rewritten as:

$$\Psi=A' \Sigma_{12} B-\frac{1}{2}\lambda(A' \Sigma_{11} A-1)-\frac{1}{2}\mu(B' \Sigma_{22} B-1)$$

If we make $\frac{\partial \Psi}{\partial A}=0$ and $\frac{\partial \Psi}{\partial B}=0$, we get

$$\begin{pmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{pmatrix}\begin{pmatrix} A \\ B \end{pmatrix}= 0$$

It must satisfy that

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix}= 0 \quad (1)$$

We can prove that the equation (1) has P roots, $\lambda_1, \lambda_2 ..... \lambda_p$ [2]. To solve equation (1), the canonical correlation problem is transformed into a general characteristic value problem. The characteristic vectors $(a^{(1)},b^{(1)})$, $(a^{(2)},b^{(2)})$...., $(a^{(p)},b^{(p)})$ corresponding to $\lambda_1, \lambda_2 ....., \lambda_p$ are the row vectors of transformation Matrixes A and B. Finally we can map the test vectors into training space by calculating $B^{-1}A(X^{(2)}- \overline{X}^{(2)})+\overline{X}^{(1)}$. Dong Yu find that retraining with transformed speech has the best compensating effect[1]. But when we considered the on-line application of this technique, we did not retrain model and only transformed the test cepstral vectors into the training space to recognize.

## 2.2. Approaches to Select Reference Speech and Adaptation Vocabulary

CCBC utilizes a few of training speech samples called reference speech and a few of test speech samples called adaptation speech to find the mapping between training set and test set. It is obvious that CCBC needs to know the correspondence between reference speech and adaptation speech. Because there are always many different utterances corresponding to the same word in training set, how to determine the reference utterances becomes an important issue. We have been working on three approaches to deal with this problem:

1. We can use the utterances of the speaker with the highest recognition rate in the training set.
2. We can use the utterances of the speaker in the training set who has the least acoustic distance from the test speaker. In our experiment, we used the DTW and Euclidean distance of cepstral vectors as the measurement of acoustic difference between two speakers.
3. We can use the clustering technique to find the represent utterance which is the centroid of all the utterances corresponding to the same word.

We also find the adaptation vocabulary which covers most valuable acoustic information is superior to other arbitrarily selected vocabularies. In our ASR system, the optimized acoustic model considered the diphones of INITIALs. So we chose an adaptation vocabulary which covered all the diphones of INITIALs and FINALs.

## 3. EXPERIMENTS AND RESULTS

Several experiments were performed to evaluate the recognition accuracy provided by CCBC, along with related algorithms. The database used is in Chinese and its vocabulary includes 500 isolated-words. In total there are 20 speakers, 10 of them being male speakers(m1-m10) and the other 10 speakers being female(f1-f10). The database was originally recorded by a DAT recorder and a close-talking microphone, and was sampled at 16kHz. The training set consisted of 18 speakers ( m1-m9 and f1-f9). The test set contained two speakers(m10 and f10). To establish the test set for channel changes, the speech data of three new speakers (which were two male speakers m11 and m12, and one female speaker f11) was recorded with a Creative 16-bit Sound Blaster and the associated microphone, utilizing the same vocabulary. To introduce the real noisy environment, we recorded a background noise of 80dB in a noisy market at first. Then we recorded the speech data of two male speakers(m13 and m14) in this background noise with the Sound Blaster.

In recognition, each utterance was represented in parametric form by computing its 12 MFCCs and 12 delta MFCCs at a rate of 12ms with a Hamming window of 24ms. However, only the 12 MFCCs were used in the adaptation procedure of CCBC, and the delta values of those transformed vector coefficients were computed during the recognition procedure. The recognizer used here is a basic VQ/DHMM.

Totally we performed four experiments under five mismatch conditions. The results of these experiment are listed in Table 1~7.

| Speaker | No adaptation | CCBC with S1 | CCBC with S2 | CCBC with S3 |
|---|---|---|---|---|
| m10 | 9.4 | 7.6 | 6.8 | 6.6 |
| f10 | 10.0 | 5.4 | 5.2 | 5.2 |

**Table 1:** Percentage error rate for speaker adaptation using CCBC (m10 and f10). S1, S2 and S3 represent three approaches to select reference speech.

| Adaptation technique | m11 | m12 | f11 |
|---|---|---|---|
| No Adaptation | 22.0 | 20.2 | 27.4 |
| CCBC with S1 | 9.8 | 11.6 | 7.6 |
| CCBC with S2 | 10.8 | 10.6 | 7.0 |
| CCBC with S3 | 7.8 | 8.6 | 7.2 |
| Cepstral normalization | 13.4 | 13.2 | 14.6 |
| RASTA | 16.4 | 14.6 | 17.6 |

**Table 2:** Comparison of percentage error rate for speaker and channel adaptation using CCBC, cepstral normalization and RASTA cepstral processing.

| Adaptation technique | SNR(dB) | | | | |
|---|---|---|---|---|---|
| | 40 | 30 | 20 | 10 | 0 |
| No adaptation | 9.4 | 10 | 18.2 | 34.2 | 44.6 |
| CCBC with S3 | 6.6 | 7.2 | 9.8 | 14.8 | 22.1 |
| Lin-Log RASTA | 13.6 | 14.6 | 15 | 22.4 | 40.9 |

**Table 3:** Comparison of percentage error rate for speaker and noise adaptation using CCBC and Lin-Log RASTA, on male speaker m10 with simulated noise.

| Adaptation technique | SNR(dB) | | | | |
|---|---|---|---|---|---|
| | 40 | 30 | 20 | 10 | 0 |
| No adaptation | 10 | 11.2 | 14.2 | 35.4 | 40.2 |
| CCBC with S3 | 5.2 | 6 | 7.6 | 14 | 20 |
| Lin-Log RASTA | 13.2 | 13 | 13.6 | 27.5 | 44.3 |

**Table 4:** Comparison of percentage error rate for speaker and noise adaptation using CCBC and Lin-Log RASTA, on female speaker f10 with simulated additive noise.

| Adaptation technique | SNR(dB) | | | | |
|---|---|---|---|---|---|
| | 40 | 30 | 20 | 10 | 0 |
| No adaptation | 22 | 24.2 | 26.2 | 33.6 | 56.6 |
| CCBC with S3 | 7.8 | 9.2 | 9.2 | 10 | 21.2 |
| Lin-Log RASTA | 22.7 | 23 | 17.8 | 26.6 | 50.3 |

**Table 5:** Comparison of percentage error rate for speaker, channel and noise adaptation using CCBC and Lin-Log RASTA, on male speaker m11 with simulated additive noise.

| Adaptation technique | SNR(dB) | | | | |
|---|---|---|---|---|---|
| | 40 | 30 | 20 | 10 | 0 |
| No adaptation | 27.4 | 27.6 | 32.8 | 39.2 | 81.8 |
| CCBC with S3 | 7.2 | 7.6 | 8.4 | 10.6 | 19.4 |
| Lin-Log RASTA | 24.5 | 25.1 | 22.4 | 26.8 | 58.2 |

**Table 6:** Comparison of percentage error rate for speaker, channel and noise adaptation using CCBC and Lin-Log RASTA, on female speaker f11 with simulated additive noise.

| Adaptation technique | m13 | m14 |
|---|---|---|
| No adaptation | 23.6 | 33.2 |
| CCBC with S1 | 15.0 | 12.2 |
| CCBC with S2 | 19.4 | 13.8 |
| CCBC with S3 | 14.8 | 11.2 |
| Lin-log RASTA | 15.6 | 17.8 |

**Table 7:** Comparison of Percentage error for speaker, channel and noise adaptation using CCBC and Lin-Log RASTA, on male speaker m13 and m14 with 80dB background noise recorded in a market.

In the first experiment we examined the performance of CCBC in dealing with mismatch condition 1: different speaker, by test on speaker m10 and f10(Table 1). In the second experiment we compensated mismatch condition 2: different speaker and channel, in case of speaker m11, m12 and f11(Table 2). We also compared the recognition accuracy obtained by using the CCBC with that of using two well-used compensation technique: (1) cepstral mean normalization, (2) RASTA cepstral processing. In the third experiment, we added additive Gaussian white noise into the speech signal used in the first experiment(m10 and f10). This simulated the third mismatch condition including the difference of speaker and the effects of noisy environment. Besides CCBC, Lin-Log RASTA spectral processing was also used to deal with this case(Table 3 and 4). In the fourth experiment, we utilized CCBC and Lin-Log RASTA to improve the recognition rate of the worst degraded test speech, in which all three kinds of mismatch between test set and training set were integrated. In this experiment, both simulated noisy speech(m11 and f11 with simulated additive noise) and real noisy speech(m13 and m14) was tested, which were listed as the mismatch condition 4 (Table 5 and 6) and condition 5(Table 7) respectively.

We note that CCBC well-compensated all these mismatch conditions and outperforms all other adaptation techniques. In our speaker-independent speech recognition system, the error rate of training set is 6.24%. We can see from the results that by CCBC test set can has the error rate approaching to or even better than that of the training set in most case. Only when SNR fell to be 0 dB, this desired result was not got. However, CCBC still improved the performance greatly even in low SNR case. The general decrease of error rate is two times. In a extreme case(m11), the de-

crease of error rate reaches four times. We compared three kinds of approaches to select reference speech data on three mismatch conditions. Except on speaker f11, the scheme 3 has the lowest error rate on all other cases. While scheme 1 and 2 select the utterances only spoken by one of the speaker in the training set, scheme 3 select the class centroid of all the utterances corresponding to the same word. Compared with the other two schemes, it can embody the common acoustic characteristic of training set. So the reference speech chosen by scheme 3 can better match the acoustic model than the reference speech chosen by the other schemes.

Although cepstral normalization has a better performance than the RASTA algorithm, its error rates are always approximately two-fold of that of CCBC. This is because cepstral normalizaion only compensate the shift of cepstral means and CCBC can compensate both the shift of cepstral means and norm shrinks. RASTA removes the slow variation in speech signal, but it may also remove some useful low-frequency speech component. So it did not work as well as CCBC and cepstral normalization. Lin-Log RASTA has positive effect on improving speech recognition rate in the worst mismatch conditions because it can compensate both additive noise and convolutional noise. However, its performance is worse than CCBC because it may be viewed as a form of noise-masking[9], which can only make the feature space insensitive to noise and do not compensate the spectral difference between distortion speech and clean speech. Lin-Log RASTA needs to retrain model according to different noise level, which is hard to apply in real application.

## 4. CONCLUSIONS

Although there exists the problem to select appropriate reference speech, the proposed CCBC algorithm can make our speaker-independent speech recognition system robust to all three kinds of mismatch between training set and test set. CCBC provided significant improvement in performance on five tested mismatch conditions. In most case, the recognition rate provided by CCBC can approach to that of the training set.

Compared with other adaptation techniques used in this paper, CCBC not only has the best compensated effect, but also is most suitable to be an on-line adaptation technique. Its calculation procedure is definite and highly efficient. We can only transform the tested speech to training space without retraining the model. It do not need to know any knowledge of distortion sources and noisy level.

Another attractive feature of CCBC is that it can be combined with other adaptation techniques. We had tried replacing the cepstral mean value of reference speech with that of calculated by cepstral normalization. This improved the performance of CCBC. The Lin-Log RASTA needs to be retrained according to different noise level. If we combine it with CCBC, we can solve this problem by mapping the spectrum obtained from a J value( in the logarithmic transform of Lin-Log RASTA) corresponding to the noise level of test speech to a spectrum processed with a J value for clean speech. Thus we only need to train acoustic models in clean speech.

## 5. REFERENCES


1. Dong Yu and Taiyi Huang, “Canonical Correlation Based Compensation Approach for Robust Speech Recognition in Noisy Environment”, EUROSPEECH 95, pp477-480.

2. “An Introduction to Multivariate Statistical Analysis”, T.W.Anderson, 2nd Edition, 1984.

3. Hynek Hermansky and Nelson Morgan, “RASTA Processing of Speech”, IEEE Transactions on Speech and Audio Processing, Vol. 2., No.4, October 1994.

4. Joachim Koehler and Nelson Morgan etc., “ Integrating RASTA-PLP into Speech Recognition”, ICASSP 1994, I421-I424.

5. Yunxin Zhao, “Self-learning speaker and channel adaptation based on spectral variation source decomposition”, Speech Communication, April, 1995.

6. Alejandro Acero and Richard M. Stern, “ Environmental Robustness in Automatic Speech Recognition”, ICASSP 1990, pp849-852.

7. Chafic Mokbel and Gerard Chollet, “Word Recognition in the Car( Speech Enhancement/ Spectral Transformations)”, ICASSP, 1991, pp925-929.

8. Jane Chang and Victor Zue, “A Study of Speech Recognition System Robustness to Microphone Variations: Experiments in Phonetic Classification”, ICSLP 94.

9. J.P. Openshaw and J.S. Manson, “On the Limitations of Cepstral features in Noise”, ICASSP 1994, II49-II53.

10. H.C. Choi and R.W. King, “Speaker Adaptation through Spectral Transformation for HMM based Speech Recognition”, 1994 International Symposium on Speech, Image Processing and Neural Networks, pp686-689.

11. W. Van Summers and David B. Pisoni etc. “Effects of Noise on Speech Production: Acoustic and Perceptual analyses”, J. Acoust. Soc. Am., September 1988, pp917-pp926.

12. Aaron E. Rosenberg, Chin-Hui Lee, Frank K. Soong, “Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification”, ICSLIP 1994