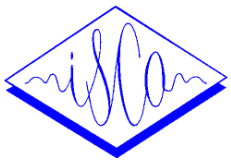

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

CANONICAL CORRELATION BASED COMPENSATION APPROACH FOR ROBUST SPEECH RECOGNITION IN NOISY ENVIRONMENT

Dong Yu, Taiyi Huang
email: huang@prldec3.ia.ac.cn
National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
P.O. Box 2728, Beijing 100080
P.R. China

ABSTRACT

In this paper we propose a novel compensation approach named Canonical Correlation Based Compensation(CCBC) to improve the performance of recognizers under noisy environment. In practical speech recognition applications, the mismatching between training and testing environments often seriously diminish recognition accuracy. Because the testing environments are not always known beforehand, The adaptive compensation framework is a practical approach to cope with this problem. While other compensation methods often deal with only some of the cepstrum changes, and is effective only for specific condition, the new approach proposed here is noise independent and can compensate all of the three main differences between two environments, i.e. mean value shift, norm shrink and the bad correlation of each dimension between training and testing speech. The experimental results show that our method has very good compensation effect.

1. INTRODUCTION

The performance levels of most current speech recognizers degrade significantly when training and testing environments are different. These environmental differences can be seen as two types of noises: the additive noise and multiplicative noise in the spectral domain. Actually, compared with reference speech, the cepstrum of "noisy" speech has three main changes: mean value shift, norm shrink and more important, the correlation of each dimension between reference speech and noisy speech of the same utterance becomes bad. If we could find a mapping to compensate such variation then we can improve the performance effectively. Since the environmental noise may not be known in advance, the idea of adaptive compensation is practical.

Several methods has been proposed to find the compensation mapping[1-4]. Though these methods are successful they are not good enough, because most of them have not dealt with the first and/or third changes in noisy cepstrum. In addition, most of these methods are noise specific methods. They dependent on the type of noise.

In this paper we propose a novel compensation approach named Canonical Correlation Based Compensation(CCBC) to improve the performance of recognizers under noisy environment. The main idea is to find two mappings to change cepstrum in training space and testing space into the same reference space called canonical correlation space. At such a reference space, the transformed training data is consistent with transformed testing data and so training and testing in the same reference space is better than training and testing in different spaces. This novel approach does not need the assumption that the noise is independent with speech and can be used to compensate any type of variation between training and testing environments (such as difference caused by different emotional condition, speaker, background, etc.). Furthermore, this novel approach can compensate all three mismatches mentioned above.

In the second part we will introduce the canonical correlation. In the section three we will introduce the procedure used in our system to find the compensation mapping. The experimental results are listed in the forth part. At the end of the paper is a discussion.

2. CANONICAL CORRELATION

Canonical Correlation is originally used to analyze the correlations between the variables of two sets which has a joint distribution[5]. The purpose of this analysis is to find a new coordinates which display unambiguously the system of correlation, i.e, we find linear combinations of variables in the sets that have maximum correlation, these linear combinations are the first coordinates in the new system; then a second linear combinations in each set is sought such that the correlation between these is the maximum and this second combinations is uncorrelated with the first linear combinations. The procedure is continued until the two coordinate systems are completely specified.

In other words, suppose the random vectors $x^{(1)}$ and $x^{(2)}$ from sets $X^{(1)}$ and $X^{(2)}$ both of p components has the covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} \Sigma_{12} \\ \Sigma_{21} \Sigma_{22} \end{pmatrix}$$

The canonical correlation problem is to find matrices (or mappings) $A^{(1)}$ and $A^{(2)}$ such that $y^{(1)} = A^{(1)}x^{(1)}$ and $y^{(2)} = A^{(2)}x^{(2)}$ has the relationship:

for every $i = 1,2,.....p-1$

$$\lambda_i = \max_{a_i^{(1)}, a_i^{(2)}} E(a_i^{(1)T}x^{(1)})(a_i^{(2)T}x^{(2)})^T$$

$$E\left(a_i^{(1)T}x^{(1)}\right)^2 = E\left(a_i^{(2)T}x^{(2)}\right)^2 = 1$$

$$E(a_i^{(1)T}x^{(1)})(a_j^{(1)T}x^{(1)})^T = 0 \qquad j = 1,2,.....i-1$$

$$E(a_i^{(2)T}x^{(2)})(a_j^{(2)T}x^{(2)})^T = 0 \qquad j = 1,2,.....i-1$$

where λ_i is the ith correlation in the reference (or new) coordinate system, $E(u)$ is the expectation of u and $a_i^{(k)}$ is the ith row vector of transformation matrix $A^{(k)}$, i.e.

$$A^{(k)} = (a_1^{(k)}, a_2^{(k)}, ...a_p^{(k)})^T$$

It can be proved[5] that the canonical correlations $\lambda_1, \lambda_2 ....., \lambda_{p_1}$ satisfy

$$\left|\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \lambda^2\Sigma_{11}\right| = 0$$

That is, the canonical correlation problem can be transformed to a general characteristic value problem and can be solved with many methods. The characteristic vectors is corresponding to the row vectors of mappings $A^{(1)}$ and $A^{(2)}$.

3. CANONICAL CORRELATION BASED COMPENSATION

The performance levels of most current speech recognizers degrade significantly when training and testing environments are different. The performance degradation is mainly due to the mismatching between training and testing environment. Mansour and Juang [4] observed that when corrupted with white noise, the Cepstral feature vectors will have two changes: norm shrink and relative robustness of the Cepstral vector orientation. During the study of noisy speech recognition, we found that besides the norm shrink, the noisy cepstrums are undertaken other two mismatching: mean value shift and inconsistent of each dimension between training speech and testing speech.

If we regard training set and testing set are the sets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ mentioned in the section 2, then we can found that, in general, the relationship between one dimension of variables in training space and testing space can be classified as four types (shown in Fig. 1). Fig. 1(a) is the worst condition where the relationship between $x_i^{(1)}$ and $x_i^{(2)}$ is totally randomed, i.e., the pair $(x_i^{(1)}, x_i^{(2)})$ is scattered in the whole plane. In fact, in such a condition, the speech is undertaken all the three changes mentioned before. Fig. 1(b) is a condition with norm shrink and bias. Fig 1.(c) which has only bias is better than Fig. 1(a) and Fig. 1(b). Fig 1(d) is the condition we want to pursue. It is clearly that, in the Fig. 1(d) condition, the model trained in $X^{(1)}$ space can be used to test the speech in $X^{(2)}$ space. Canonical Correlation method is an approach to transform training space and testing space to condition Fig. 1(d).

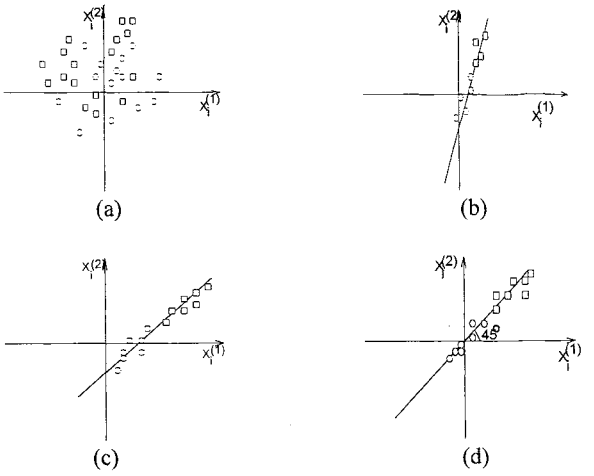

Fig. 1 relationship between two sets of variables

The procedure of CCBC can be divided into several steps:

1. Estimate the mapping. Several utterances are enough for finding this mapping;

   a. find the matching pair of cepstrum between two environment using DTW;

   b. estimate transformation matrix $A^{(1)}$ and $A^{(2)}$ to make $y^{(1)} = A^{(1)}x^{(1)}$ and $y^{(2)} = A^{(2)}x^{(2)}$ so that $y^{(1)}$ and $y^{(2)}$ have maximum correlation, where $x^{(1)}$ is the cepstrum vectors of training speech and $x^{(2)}$ is the cepstrum vectors of testing speech. $A^{(1)}$ and $A^{(2)}$ are canonical matrix to make $y^{(1)}$ and $y^{(2)}$ has same norm which compensate the norm shrink of noisy cepstrum;

   c. estimate mean value $\bar{x}^{(1)} = E(x^{(1)})$ and $\bar{x}^{(2)} = E(x^{(2)})$;

2. Transform training speech and testing speech to same space.

   condition A: transformed to training space.

   $z^{(1)} = x^{(1)}$

$$z^{(2)} = A^{(1)^{-1}} A^{(2)} (x^{(2)} - \bar{x}^{(2)}) + x^{(1)}$$

condition B: transformed to testing space.

$$z^{(1)} = A^{(2)^{-1}} A^{(1)} (x^{(1)} - \bar{x}^{(1)}) + x^{(2)}$$
$$z^{(2)} = x^{(2)}$$

condition C: transformed to canonical correlation reference space.

$$z^{(1)} = A^{(1)} \left(x^{(1)} - \bar{x}^{(1)} \right)$$
$$z^{(2)} = A^{(2)} \left(x^{(2)} - \bar{x}^{(2)} \right)$$

By these transformation, the three mismatches mentioned above have all been compensated.

3. Retrain recognizer using transformed speech $z^{(k)}$ or find the transformed model directly from original model.

condition B: transformed to testing space.

$$\mu'' = A^{(1)^{-1}} A^{(2)} (\mu - \bar{x}^{(2)}) + \bar{x}^{(1)}$$
$$\sigma'' = (A^{(1)^{-1}} A^{(2)}) * \sigma * (A^{(1)^{-1}} A^{(2)})^T$$

condition C: transformed to canonical correlation reference space.

$$\mu'' = A(\mu - \bar{x})$$
$$\sigma'' = A * \sigma * A^T$$

where σ and μ are the covariance and mean value of old model and σ'' and μ'' are the covariance and mean value of new model.

4. using new model to submit a test.

Note That the utterances used to find the mapping are not the same as the utterances used as testing set.

4. EXPERIMENT

To evaluate the power of this method in noisy speech recognition, we have done experiments on initials /b/, /d/ and /g/. We have three groups of data each has 55 /b/, 62 /d/ and 57 /g/ and use one group to train the original model, another group to test the new model. The rest group is used to estimate the mapping matrix. In our experiment we use 15 utterances (5 for each initial) to estimate the transformation matrices. The noise used here is additive Gaussian white noise and is added using a program. The recognizer used here is the basic CDHMM. Because the ultimate space can either be the training space, testing space or canonical correlation reference space, we submitted results in all these conditions. To evaluate our method, we compare CCBC's performance with that of both R1, i.e. training with clean speech and testing with noisy speech and R2, training with noisy speech and testing with noisy speech. The results are listed in Table 1. 2 and 3. where

condition R1: training with clean speech and testing with noisy speech;
condition R2: training with noisy speech and testing with noisy speech;
condition A: transformed to training space, training with clean speech and testing with transformed speech;
condition B1: transformed to testing space, training with transformed speech and testing with testing speech, the new model is derived directly from old model;
condition B2: transformed to testing space, training with transformed speech and testing with testing speech, the new model is retrained with transformed data.
condition C1: transformed to reference space, training with transformed speech and testing with transformed speech, the new model is derived directly from old model;
condition C2: transformed to reference space, training with transformed speech and testing with transformed speech, the new model is retrained from transformed data;

Cond.	SNR				
	∞ dB	30 dB	20 dB	10 dB	0 dB
R1	87.93%	87.35%	85.31%	66.27%	41.57%
R2	87.93%	87.93%	88.51%	85.06%	67.24%
A	87.93%	85.06%	83.33%	75.86%	66.09%

Table 1. White Noise Case 1: transformed to training space

Cond.	SNR				
	∞ dB	30 dB	20 dB	10 dB	0 dB
R1	87.93%	87.35%	85.31%	66.27%	41.57%
R2	87.93%	87.93%	88.51%	85.06%	67.24%
B1	87.35%	82.76%	81.03%	74.71%	64.94%
B2	87.93%	85.06%	83.91%	77.01%	67.24%

Table 2. White Noise Case 2: transformed to testing space

Cond.	SNR				
	∞ dB	30 dB	20 dB	10 dB	0 dB
R1	87.93%	87.35%	85.31%	66.27%	41.57%
R2	87.93%	87.93%	88.51%	85.06%	67.24%
C1	87.93%	85.06%	84.48%	78.74%	68.39%
C2	87.93%	85.63%	87.93%	81.03%	71.26%

Table 3. White Noise Case 3: transformed to canonical correlation reference space

These results show that the new method proposed here can improve the performance greatly. First, compared with condition R1, i.e. training with clean speech and testing with noisy speech, in the noisy environment ($SNR \leq 10dB$), whatever space the training and testing space changed to, the

performance is better than that of condition R1. Second, the performance of using this method is even better than that of condition R2, i.e. training with noisy speech and testing with noisy speech.

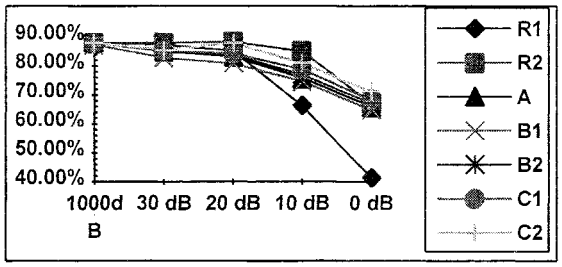

Fig. 2. White Noise Case: Comparison results of seven conditions

Another observation is that retraining the new model is better than deriving the new model directly from old model and transformed to the canonical correlation space can obtain the best result.

To show that the method proposed here can improve the performance of recognizer under other noises we have done an experiment on speaker-independent word based continuous speech recognition system. The training set is composed of 500*18 words pronounced by 18 people and recorded with a digital recorder and TMS320C30 in a quiet room. The testing set is composed of 500 utterances spoken by the 19th person and recorded with Sound Blaster in normal office. The comparison results are listed in Table 4 where the condition R1 and A are the same as what indicated above. We can conclude from Table 4 that the noval approach can cope with these unclear noises caused by different speaker, different channel and different environment.

condition	Top 1	Top 2	Top 3	Top 4	Top 5
R1	87.4%	94.0%	95.6%	97.2%	97.8%
A	94.0%	97.6%	98.6%	99.2%	99.4%

Table 4. Different Speaker and Channel Case

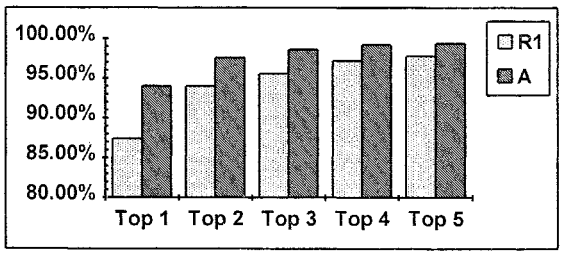

Fig. 3. Different Speaker and Channel Case:Comparison results

5. CONCLUSION AND DISCUSSION

The method proposed in this paper is based on canonical correlation analysis. The main idea of this method is to find a reference space which training space and testing space can be easily transformed into using linear mapping. In such a space, transformed training data and testing data can be consistent and all three mismatches mentioned in the paper can be compensated. This approach is independent on the source of variation in speech and need not the assumption that the noise is independent with speech, so it can be used to deal with all variations in speech such as variation caused by different speakers, emotional conditions and channels. Furthermore, this method can be used with other method such as Spectrum Subtraction and RASTA-PLP to better the result further.

One of the most attractive feature of CCBC method is the self-organization capability. We can learn from the experiment that, when SNR becomes 0dB, the performance of condition B2, C1 and C2 is better than that of condition R2. At the first glance, the performance of condition R2 should be the best, because in this condition, the training data and testing data are the same and so have not the mismatching problem. But we know that, when corrupted with adverse noise, the speech will be greatly distorted and the difference between each pair of classes will be blurred. The result of these effects is that it will diminish the performance of classifiers. The CCBC method has a good feature. As we can see in the section two, this method can decrease the space dimension by only retain the most important dimensions. This feature makes the CCBC method a chance to take advantage of the information in reference clean speech to improve the performance of seriously corrupted speech.

6. REFERENCES

[1] Lee-Min Lee and Hsiao-Chuan Wang, "A Study of Cepstral and Delta Cepstral Coefficients for Noisy Speech Recognition", Proceeding ICSLP 94;

[2] Keizaburo TAKAGI, Hiroaki HATTORI and Takao WATANABE, "Speech Recognition with Rapid Environment Adaptation by Spectrum Equalization", Proceeding ICSLP 94;

[3] William C. Treurniet and Yifan Gong, "Noise Independent Speech Recognition for a Variety of Noise Types", Proceeding ICASSP 94;

[4] D.Mansour and B.H. Juang. A family of distortion measures based upon projection operation for robust speech recognition. IEEE Trans. ASSP, Vol.37(No.11):pp.1659-1671. November 1989.

[5] "An Introduction to Multivariate statistical Analysis", T.W. Anderson, 2nd Edition, 1984.