

CHANNEL AND NOISE NORMALIZATION USING AFFINE TRANSFORMED CEPSTRUM

Xiaoyu Zhang and Richard J. Mammone

T-NETIX Inc., Denver, Colorado

xiaoyu@tnetix.com

ABSTRACT

This paper addresses the environmental mismatch problem that arises from noise and channel variabilities. A new feature mapping technique based on an optimal affine transform of the cepstrum is proposed to solve the mismatch problem observed over the speaker recognition systems. It is designed based on the fact that both the channel and noise interferences basically cause the cepstrum space to undergo an affine transformation. By taking an inverse transformation, we can easily decouple from the speech the effects of the channel and noise. Alternatively, we can take a forward transform of the training data to simulate the operating conditions.

1. INTRODUCTION

The current speaker and speech recognition technologies have enabled a computer to successfully perform voice verification and speaker identification in noise-free environment. However, these automatic recognizers are confronted with the challenges emerging from the customized applications where there is generally a mismatch condition between the training environment and the diverse and unpredictable operating environments. For example, in the telephone-based voice verification system, a customer first calls in over a regular electret phone to enroll his voice prints into the database. A few minutes later, after the system has been trained to recognize his voice, the customer dials in again and wants to have himself be verified by the system. However, he calls over a cellular phone this time. It is very likely that the system will not be able to recognize him simply because he is using a cellular phone rather than the electret handset he used for enrollment. The problem here is the mismatched channel between the electret phone and the cellular phone. Channel mismatch can also result from the instability of the circuits, the change of the central switching office, and the switch of the service zones in the cellular network, etc. Other than channel mismatch, noise from the acoustical background and the speaker variability due to the background noise are also the causes of the mismatch conditions. It has been found in practice that the mismatch conditions generally cause a much more dramatic

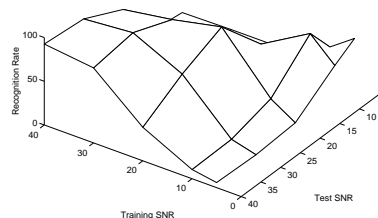


Figure 1: Recognition rate under various combinations of training and test SNR.

degradation of performance than the channel and noise variations on their own. This can be made clear by the experiments conducted on a speaker identification system. The recognition rates obtained when the system is tested in different training and test SNR conditions are plotted in Fig. . The x axis represents the SNR of the training data and the SNR of the test data is represented by y . We can see that the performance of the system is somehow maintained when the training and testing SNRs are matched (as shown by points along the diagonal line in the plot), as compared to the off-diagonal points which represent the situations when the training SNR is not matched with the test SNR.

A family of adaptation methods that target at establishing a mapping function $y = f(x)$ between the training data x and the test data y have been proposed to attack the mismatch problem in [1][2][3][4][5][6].

In this paper, we propose using affine transform $y = Ax + b$ to model the channel and noise mismatch. Fig. 2 shows the scatter plots of the cepstral coefficients we generated by passing the speech through the wireline simulator and/or corrupting the speech with white Gaussian noise. If we pass the speech through the wireline simulator, the cepstral clusters representing different sounds are shifted in the same direction regardless of what the sounds are (Fig. 2(a)). If the speech is corrupted by white Gaussian noise of 15dB SNR, the cepstral clusters are rescaled

and all of them move towards the origin(Fig. 2(b)). In the case that the speech is both filtered by the channel and corrupted by noise, the translation and the rescaling take their orders depending on whether the speech is first passed through the wireline simulator and then corrupted by noise (Fig. 2(c)), or vice versa (Fig. 2(d)). The affine transform is a generalized representation of these variations in the feature space. The linear part \mathbf{A} represents precisely the effects of noise and the offset \mathbf{b} represents the degradation due to the channel.

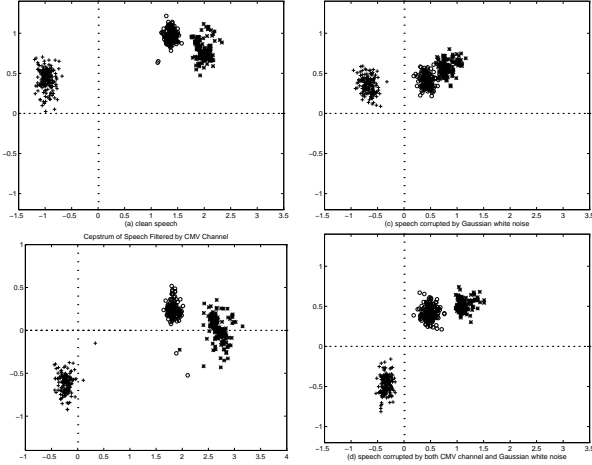


Figure 2: The movement of cepstral clusters under various conditions. 'a' is for the sound /a/, 'o' for /n/, and '+' for /sh/. (a) Cepstrum of the original clean speech; (b) Cepstrum of signals filtered by continental U.S. mid voice channel (CMV); (c) Cepstrum of the speech with 15 dB SNR. The noise is additive and white Gaussian (AWG); (d) Cepstrum of the speech filtered by the CMV channel and then corrupted by AWG noise of 15 dB SNR.

2. AFFINE TRANSFORMED CEPSTRUM

We have discussed in an earlier paper [7][8] that when the speech is interfered by the linear channel and random noise, both the linear predictor coefficients and the cepstral coefficients experience an affine transform in the feature domain. In this section, we will briefly review these derivations.

2.1 Affine Transform of Predictor Coefficients

First, we study the affine transform of the coefficients of the linear prediction model by examining noise and the channel interferences separately.

Noise coming from the acoustical background are generally simplified as additive white random noise. The

speech signal corrupted by the random noise is therefore represented as

$$s'(n) = s(n) + e(n), \quad (1)$$

where $e(n)$ denotes the noise, and

$$E[e(n)] = 0, \quad E[e^2(n)] = \sigma^2, \quad E[e(m)e(n)] = 0 \quad (m \neq n). \quad (2)$$

It is zero mean and uncorrelated.

By studying the autocorrelation of the signals obtained before and after the noise corruption, we found that the predictor coefficients of the noise-corrupted speech $s'(n)$ can be represented as

$$\begin{aligned} \mathbf{a}' &= \mathbf{R}_{s'}^{-1} \mathbf{r}_{s'} \\ &= (\mathbf{R}_s + \sigma^2 \mathbf{I})^{-1} \mathbf{r}_s = (\mathbf{R}_s + \sigma^2 \mathbf{I})^{-1} \mathbf{R}_s \mathbf{a}, \end{aligned} \quad (3)$$

where the relation between the autocorrelation $\mathbf{R}_{s'}$ (matrix), $\mathbf{r}_{s'}$ (vector) of the noise-corrupted speech and the autocorrelation \mathbf{R}_s , \mathbf{r}_s of the input clean speech can be formulated as

$$\mathbf{R}_{s'} = \mathbf{R}_s + \sigma^2 \mathbf{I}, \quad \text{and} \quad \mathbf{r}_{s'} = \mathbf{r}_s. \quad (4)$$

When a speech signal is passed through a linear channel, the filtered signal $s'(n)$ obtained at the output of the channel is

$$s'(n) = s(n) \otimes h(n), \quad (5)$$

where $h(n)$ is the impulse response of the channel. The autocorrelation of the output signal $s'(n)$ can be represented by

$$r_{s'}(k) = [h(n) \otimes h(-n)] \otimes r_s(k) = r_h(k) \otimes r_s(k), \quad (6)$$

where $r_s(k)$ is the autocorrelation of the input $s(n)$.

The predictor coefficients of the output $s'(n)$ is a linear translation of the predictor coefficients of the input $s(n)$ given by

$$\begin{aligned} \mathbf{a}_{s'} &= \mathbf{R}_{s'}^{-1} \mathbf{r}_{s'} \\ &= [(\mathbf{HS})^T \mathbf{HS}]^{-1} (\mathbf{HS})^T \hat{\mathbf{H}} \hat{\mathbf{s}} = (\mathbf{HS})^{-1} \hat{\mathbf{H}} \hat{\mathbf{s}} \\ &= (\mathbf{HS})^{-1} \begin{pmatrix} \mathbf{h}_1 & \mathbf{H} \end{pmatrix} \begin{pmatrix} s(0) \\ \mathbf{s} \end{pmatrix} \\ &= (\mathbf{HS})^{-1} (s(0) \mathbf{h}_1 + \mathbf{H} \mathbf{s}) \\ &= (\mathbf{HS})^{-1} (s(0) \mathbf{h}_1 + \mathbf{H} \mathbf{s}) \\ &= s(0) (\mathbf{HS})^{-1} \mathbf{h}_1 + \mathbf{a}, \end{aligned} \quad (7)$$

where $\mathbf{h}_1 = [h(1), h(2), \dots, h(N)]^T$ and $h(n)$ is the impulse response of the channel. The matrix \mathbf{H} can be derived from $h(n)$ and the matrix \mathbf{H}_s can be derived from the impulse response of the linear predictive function. The matrix \mathbf{S} is dependent on the input signal $s(n)$ [7][8]. Therefore, the offset term is a function of the input and the impulse response of the linear channel.

2.2 Affine Transform of Cepstral Coefficients

The cepstrum is by definition

$$c(n) = \mathcal{Z}^{-1}[\log S(z)] = \mathcal{Z}^{-1}\left[\log \frac{1}{1 - \sum_{i=1}^p a_i z^{-i}}\right]. \quad (8)$$

We can prove that the cepstrum is a function of the predictor coefficients and the impulse response. It is given by

$$c(n) = \frac{1}{n} \sum_{i=1}^p i a_i h_s(n-i). \quad (9)$$

The matrix form of this equation is

$$\begin{aligned} \begin{pmatrix} c(1) \\ c(2) \\ \vdots \\ c(N) \end{pmatrix} &= \begin{pmatrix} 1 & & & \\ & 1/2 & & \\ & & \ddots & \\ & & & 1/N \end{pmatrix} \\ &\times \begin{pmatrix} h_s(0) & 0 & \dots & 0 \\ h_s(1) & h_s(0) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_s(p-1) & h_s(p-2) & \dots & h_s(0) \\ \vdots & \vdots & \ddots & \vdots \\ h_s(N-1) & h_s(N-2) & \dots & h_s(N-p) \end{pmatrix} \\ &\times \begin{pmatrix} 1 & & & \\ & 2 & & \\ & & \ddots & \\ & & & p \end{pmatrix} \begin{pmatrix} a(1) \\ a(2) \\ \vdots \\ a(p) \end{pmatrix} \\ &= \mathbf{D}_1 \mathbf{H}_s \mathbf{D}_2 \mathbf{a}. \end{aligned} \quad (10)$$

Therefore, when the speech is passed through a linear channel, a set of degraded cepstral vectors can be obtained by

$$\mathbf{c}' = \mathbf{D}_1 \mathbf{H}_s' \mathbf{D}_2 \mathbf{a}' = \mathbf{D}_1 \mathbf{H} \mathbf{H}_s \mathbf{D}_2 \mathbf{a}'. \quad (11)$$

Since the predictor coefficients of the filtered speech is a shifted version of the original one, we have

$$\begin{aligned} \mathbf{c}' &= \mathbf{D}_1 \mathbf{H} \mathbf{H}_s \mathbf{D}_2 [\mathbf{a} + s(0)(\mathbf{H}\mathbf{S})^{-1} \mathbf{h}_1] \\ &= \mathbf{D}_1 \mathbf{H} \mathbf{D}_1^{-1} \mathbf{c} + s(0) \mathbf{D}_1 \mathbf{H} \mathbf{H}_s \mathbf{D}_2 (\mathbf{H}\mathbf{S})^{-1} \mathbf{h}_1. \end{aligned} \quad (12)$$

The cepstral coefficients of the speech are translated when the speech is filtered by a linear channel.

In the case that the speech is contaminated by additive, white noise, the cepstral coefficients becomes

$$\begin{aligned} \mathbf{c}' &= \mathbf{D}_1 \mathbf{H}_s' \mathbf{D}_2 \mathbf{a}' \\ &= \mathbf{D}_1 \mathbf{H}_s \mathbf{D}_2 (\mathbf{R}_s + \sigma^2 \mathbf{I})^{-1} \mathbf{R}_s \mathbf{a} \\ &= \mathbf{D}_1 \mathbf{H}_s \mathbf{D}_2 (\mathbf{R}_s + \sigma^2 \mathbf{I})^{-1} \mathbf{R}_s (\mathbf{D}_1 \mathbf{H}_s \mathbf{D}_2)^{-1} \mathbf{a}. \end{aligned} \quad (13)$$

The degraded cepstral vectors are obtained by multiplying the original vectors with a sound-dependent matrix. It is equivalent to taking a linear transformation of the original cepstral vectors.

A speech signal is usually passed through multiple sources of degradation before it arrives at the receiver's end. For instance, the speech may be contaminated by additive noise from the background, and then distorted by the channel embedded in a cellular network, or the nonlinear output .vs. input response of a carbon button handset. In the case that the speech is first corrupted by additive and then filtered by a linear channel, the cepstrum actually goes through two affine transforms, i.e.,

$$\mathbf{c}_{observed} = \mathbf{c}_{noisy} + \mathbf{b} = \mathbf{A} \mathbf{c}_{original} + \mathbf{b}. \quad (14)$$

If we let the speech be filtered by the channel first, and then be passed through a noisy environment, the transformations taken on the cepstrum are switched in accordance, i.e.,

$$\mathbf{c}_{observed} = \mathbf{A} \mathbf{c}_{filtered} = \mathbf{A} \mathbf{c}_{original} + \mathbf{A} \mathbf{b}. \quad (15)$$

Therefore, the accumulated effect of the multiple degradation sources is equivalent to an affine transform.

The transformation parameters \mathbf{A} and \mathbf{b} can be solved using different methods. One of the solutions can be found by using the least square method and it is given by

$$\begin{aligned} \mathbf{a}_j &= \begin{pmatrix} a_{j1} \\ \vdots \\ a_{jq} \\ b_j \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^N \mathbf{c}_i \mathbf{c}_i^T & \sum_{i=1}^N \mathbf{c}_i \\ \left(\sum_{i=1}^N \mathbf{c}_i\right)^T & N \end{pmatrix}^{-1} \times \begin{pmatrix} \mathbf{c}_1 & \dots & \mathbf{c}_N \\ 1 & \dots & 1 \end{pmatrix} \mathbf{c}'_j \\ &\text{for } j = 1, \dots, q. \end{aligned} \quad (16)$$

\mathbf{a}_j is an augmented column vector whose first q entries form the j^{th} row of the matrix \mathbf{A} and whose last entry is the j^{th} element of the vector \mathbf{b} .

3. EXPERIMENTS

• Text-Independent Speaker Identification

The results for text-independent speaker identification on

the New England section of TIMIT database using affine transform cepstrum (ATC) are presented in table 1 and 2. Table 1 show the baseline performance for various training and test conditions. Table 2 shows the performance using ATC.

| Training speech (SNR) | Testing (SNR) and Recognition Rate (%) | | | | |
|-----------------------|--|-------|-------|-------|------|
| | clean | 30 dB | 20 dB | 10 dB | 5 dB |
| clean | 92.6 | 85.8 | 38.4 | 10.5 | 5.8 |
| 30 dB | 87.9 | 92.6 | 65.8 | 11.6 | 4.2 |
| 20 dB | 65.3 | 74.2 | 86.3 | 24.7 | 7.4 |
| 10 dB | 14.7 | 17.4 | 33.2 | 65.8 | 39.5 |
| 5 dB | 11.6 | 12.6 | 12.6 | 33.7 | 56.8 |

Table 1: Performance of the baseline speaker identification system using TIMIT database.

| Training speech (SNR) | Testing (SNR) and Recognition Rate (%) | | | | |
|-----------------------|--|-------|-------|-------|------|
| | clean | 30 dB | 20 dB | 10 dB | 5 dB |
| clean | 92.6 | 77.4 | 63.2 | 31.1 | |
| 30 dB | 83.2 | 92.6 | 72.1 | 46.8 | 26.3 |
| 20 dB | 64.2 | 75.3 | 86.0 | 58.4 | 40.5 |
| 10 dB | 32.6 | 41.1 | 48.4 | 65.3 | 45.8 |
| 5 dB | 17.9 | 28.9 | 33.7 | 46.8 | 53.7 |

Table 2: Speaker identification using the TIMIT database.

• Text-dependent Speaker Verification

In this experiment, we solve the mismatch problem by replicating the training data using the formula $\mathbf{c}_{replicated} = \mathbf{A}\mathbf{c}_{original} + \mathbf{b}$. Knowing the noise level and the spectrum characteristics of the transmission channel in the operating environment, we can take the advantage that the parameters (\mathbf{A}, \mathbf{b}) can be calculated using equation 13. This sets us free from physically collecting sample data from the operating environment. An alternative approach to computing the parameters is to use the stereo data recorded under both training and operating conditions.

The database we used in this experiment consists of 44 speakers: 27 males and 17 females. The speech is recorded over either a carbon-button phone or an electret phone. Each speaker is asked to say the word "balance beam" for 6 to 8 times. The first 3 repetitions of the word are used to train the speaker model and the rest of the repetitions are used for verification. All the utterance other than the ones used for training are used for testing. By using both the original and the replicated data for

| | Baseline | Replication |
|------------------------|----------|-------------|
| False Accept (FA) | 7.05% | 5.43% |
| False Reject (FR) | 4.73% | 3.38% |
| Equal Error Rate (EER) | 3.51% | 2.35% |

Table 3: Speaker verification on "balance beam" database using replicated data.

training, we obtained a speaker model that is much

better generalized.

4. SUMMARY

In this paper, a new method is proposed where the channel and noise variations are modeled by the affine parameters \mathbf{A} and \mathbf{b} . The speaker's data is replicated by using these transforms to extend the training data set.

One embodiment of the affine transform is to match the training environment with the operating environment by taking either a forward or an inverse transform of the cepstrum space. Another embodiment of the affine transform modeling is to extrapolate the original data by taking an affine transform $y_i = Ax_i + b$ so that it appears as though various phones and noise conditions were used for training. The transformation parameters can be calculated from the training data and the impulse response of the channel and/or the covariance of the noise. Using both x_i and y_i for training forms well generalized models.

References

- [1] A. Nádas D. Nahamoo and M.A. Picheny. Adaptive labeling: Normalization of speech by adaptive transformation based on vector quantization. *ICASSP*, pages 521–524, 1988.
- [2] L. Newneyer and M. Weintraub. Probabilistic optimum filtering for robust speech recognition. *ICASSP*, 1:417–420, 1994.
- [3] H.Gish K. Ng and J.R. Rohlicek. Robust mapping of noisy speech parameters for hmm word spotting. *ICASSP*, 2:109–112, 1992.
- [4] P.J.Moreno B.Raj E.Gouvêa and R. M. Stern. Multivariate-gaussian-based cepstral normalization for robust speech recognition. *ICASSP*, 1:137–140, May 1995.
- [5] M.G. Mazin and B.H. Juang. Signal bias removal for robust telephone based speech recognition in adverse environments. *ICASSP*, 1:445–448, May 1994.
- [6] A. Sankar and C.H. Lee. Robust speech recognition based on stochastic matching. *ICASSP*, 1:121–124, 1995.
- [7] R.J.Mammone X.Zhang and R.Ramachandran. Robust speaker recognition - a feature based approach. *IEEE Signal Processing Magazine*, July 1996.
- [8] X.Zhang and R.J.Mammone. Robust speech processing using feature space mapping. *Submitted to IEEE Trans. Speech and Audio Processing*, 1995.