PATTERN-BASED DYNAMIC COMPENSATION TOWARDS ROBUST SPEECH RECOGNITION IN MOBILE ENVIRONMENTS

Huayun Zhang   *Jun Xu*

R&D Department, InfoTalk Technology, Singapore
{huayun.zhang, jun.xu}@infotalkcorp.com

ABSTRACT

Today, the high mobility provided by wireless networks places users in a wild variety of noise and channel conditions, which poses serious challenge to telephone-base Acoustic Speech Recognition (ASR). In this paper, we propose a Pattern-based Dynamic Compensation (PDC) scheme to improve the robustness of ASR in mobile environments. In PDC, a distortion pattern-set is employed to normalize the environmental variations in training data according to a set of pre-defined application scenarios. At recognition time, instantaneous distortion is calculated as a linear combination of several possible patterns. To online estimate the combination weights robustly, a Bayesian learning process with Speech-conditioned Prior Evolution is introduced into PDC (PDC-SPE). In outdoor experiments, the PDC-SPE method outperforms other commonly used compensation/adaptation methods and leads to 20~25% relative reduction in Word Error Rate (WER) over a well-trained baseline system.

1. INTRODUCTION

The accuracy of ASR systems degrades evidently when noise/channel mismatch exists between training and testing conditions. In the case of mobile environments, the performance degradation becomes even more serious since both background and channel characteristics change every so often. The changing environment causes varying mismatch between feature domain and model distribution. Another principal source of degradation in mobile environments is the presence of coding-decoding (codec) processes in wireless links. The distortion introduced by codec can be considered as an additive non-stationary noise, which is a function of the speech signal itself. However, since most modern speech codecs are fairly complex, their effect on the original speech signal is difficult to model analytically [1].

In recent years, many studies have focused on improving ASR robustness for mismatch conditions. These methods could be divided mainly into two classes: one, to compensate the contaminated feature space before classification; second, to adapt the parameters of acoustic model to match degraded speech. In the first class, Cepstrum Mean Normalization (CMN) and RASTA filtering [2] could effectively remove channel effects with very low computational cost. For some complex methods, such as Signal Bias Removal (SBR) [3] and Stochastic Matching (SM) [4], the filtering effect is calculated more precisely through an optimal estimation with some kind of prior knowledge about speech. However, all these methods have the stability and linearity assumptions about working conditions, either explicitly or inexplicitly. Neither the time-variant factors nor the inevitable nonlinear effects in mobile environments are taken into account. The second class includes linear transform-based methods (such as MLLR [5]), Bayesian learning methods (such as Maximum A Posteriori, MAP [6]) and model composition methods (such as Parallel Model Combination, PMC [7]). Since telephone-based applications must be able to adapt to a new environment with very small amount of data, MLLR with complex parameterization can hardly achieve satisfying performance. It is even more difficult for MAP. Since MAP could only adapt those observed models, it needs relatively more data to perform well. PMC requires appropriate statistics of noise. Hence, it is inappropriate when environment characteristics are unknown or continually changing.

In our previous study [8], it is assumed that the non-speech variations due to environmental effects change slower than the variations of speech signal. Through stochastic matching between degraded speech and model distributions within a shifting short duration, instantaneous non-speech variations could be unveiled and removed from degraded feature domain. In that case, ASR accuracy is improved substantially. However, the sparse data problem due to short-time analysis makes it difficult for this method to maintain a stable performance. We have to make a trade-off carefully between reliability and effectiveness.

In this study, a distortion pattern-set is calculated in the training phase to describe the mismatch between typical application scenarios and acoustic model distributions. It is easier and more reliable to implement scenario-oriented compensation than to estimate instantaneous distortions from the ground up with very limited data. If we expect to deal with various types of possible mismatches in practice, an optimal combination across possible patterns using Bayesian learning mechanism is necessary. It is important to find appropriate prior information for attaining better performance in various cases.

In the next section, we briefly review the short-time learning scheme of [8]. Then, we focus on PDC. A Bayesian learning method is developed to calculate the combination weights in PDC. Furthermore, an incremental prior weight evolution scheme is introduced into the learning process. Section 3 is about exploring the scenario-oriented distortion patterns and associated prior weight distributions. The experiments and discussions are represented in section 4. We sum up conclusions in section 5.

2. PATTERN-BASED DYNAMIC COMPENSATION

In mobile environments, additive noise (from background) and convolutional noise (from channel) corrupt speech signal simultaneously and introduce a time-variant bias in cepstral domain. Since instantaneous distortion is a blend of various degrees of noise effect, channel effect and even the speech itself, this changing bias can be denoted as a joint function below:

$$\mathbf{b}_t = f\left(X_t, N_t, H_t\right) \tag{1}$$

1-4244-0469-X/06/$20.00 ©2006 IEEE

where X_t, H_t and N_t denote speech, filter and noise respectively.

2.1. Codebook Dependant Channel Estimation-CDCE

Through maximizing the likelihood of noisy data with respect to clean model, CDCE can calculate a changing environmental distortion. The statistics of speech is modeled by a codebook:

$$\begin{aligned} &\mathbf{\Omega}_M = \{\boldsymbol{\omega}_m\} \quad 1 \le m \le M \\ &\boldsymbol{\omega}_m = \{\alpha_{m,n}; \boldsymbol{\mu}_{m,n}; \mathbf{\Sigma}_{m,n}\} \quad 1 \le n \le N \end{aligned} \tag{2}$$

where M is the code number. Each code is a N -mixture normal distribution. $\alpha_{m,n}$, $\boldsymbol{\mu}_{m,n}$, and $\mathbf{\Sigma}_{m,n}$ represent mixture weight, mean and covariance matrix, respectively. $\mathbf{O}$ denotes the observations that fall in current analysis duration:

$$\mathbf{O} = \{\mathbf{o}_{t-T/2}, \cdots, \mathbf{o}_t, \cdots, \mathbf{o}_{t+T/2}\} \tag{3}$$

A stochastic matching between $\mathbf{O}$ and $\mathbf{\Omega}$ is conducted to maximize the likelihood:

$$\max_{\mathbf{b}_t} P(\mathbf{O}|\mathbf{\Omega}, \mathbf{b}_t) \tag{4}$$

An iterative solution for this problem could be obtained by Expectation-Maximization (EM) method:

$$\begin{aligned} &\mathbf{U}_t = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{j=t-T/2}^{j=t+T/2} \gamma_{m,n,j} \cdot \mathbf{\Sigma}_{m,n}^{-1} \\ &\mathbf{v}_t = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{j=t-T/2}^{j=t+T/2} \gamma_{m,n,j} \cdot \mathbf{\Sigma}_{m,n}^{-1}(\mathbf{o}_j - \boldsymbol{\mu}_{m,n}) \\ &\mathbf{b}_t^i = \mathbf{U}_t^{-1}\mathbf{v}_t \end{aligned} \tag{5}$$

where $\gamma_{m,n,j}$ is the occupational probability of Gaussian $\boldsymbol{\omega}_{m,n}$ at time j with the distortion assumption of previous iteration $\mathbf{b}_t^{i-1}$.

2.2 Pattern-based Dynamic Compensation (PDC)

Given the joint distribution of X_t, N_t and H_t , the distortion can be calculated as integration over the whole random space.

$$\mathbf{b}_t = \iiint_{X_t,N_t,H_t} f(X_t,N_t,H_t)\phi(X_t,N_t,H_t)dX_t dN_t dH_t \tag{6}$$

Since the number of distortion value $f(X_t,N_t,H_t)$ is not countable and the continuous joint density $\phi(X_t,N_t,H_t)$ is not available in real life, the integration has to be approximated over finite representative points:

$$\mathbf{b}_t \approx \sum_{r=1}^{R} f(X_r,N_r,H_r)P_{r,t} \tag{7}$$

A pre-calculated pattern-set is employed to describe the distortion at the typical (X_r,N_r,H_r) points for mobile applications:

$$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_R] \quad where \ \mathbf{b}_r = f(X_r,N_r,H_r) \tag{8}$$

A time-variant vector is employed to denote the instantaneous weight factors for these patterns:

$$\boldsymbol{\lambda}_t = [\lambda_{t,1}, \lambda_{t,2}, \cdots, \lambda_{t,R}]^T \quad where \ \ \lambda_{t,r} = P_{r,t} \tag{9}$$

Given the prior distribution of $\boldsymbol{\lambda}_t$, instantaneous distortion can be estimated by optimal matching between noisy data and the patterns:

$$\max_{\boldsymbol{\lambda}_t} P(\mathbf{O}|\mathbf{\Omega}_M, \mathbf{B}, \boldsymbol{\lambda}_t)P(\boldsymbol{\lambda}_t) \tag{10}$$

For simplicity, a prior normal distribution is assumed here:

$$\boldsymbol{\lambda}_t \propto N(\overline{\boldsymbol{\lambda}}, \mathbf{\Gamma}) \tag{11}$$

The iterative solution is as follows:

$$\begin{aligned} &\mathbf{U}_t = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{j=t-T/2}^{j=t+T/2} \gamma_{m,n,j} \cdot \mathbf{B}^T\mathbf{\Sigma}_{m,n}^{-1}\mathbf{B} + \tau\mathbf{\Gamma}^{-1} \\ &\mathbf{v}_t = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{j=t-T/2}^{j=t+T/2} \gamma_{m,n,j} \cdot \mathbf{\Sigma}_{m,n}^{-1}(\mathbf{o}_j - \boldsymbol{\mu}_{m,n}) + \tau\mathbf{\Gamma}^{-1}\overline{\boldsymbol{\lambda}} \\ &\boldsymbol{\lambda}_t = \mathbf{U}_t^{-1}\mathbf{v}_t \end{aligned} \tag{12}$$

where τ is introduced to adjust the contribution of prior information. When τ is set to zero, (13) becomes ML estimation. τ is adjusted by the data size available in real applications.

The probability constraints $\lambda_r \ge 0 \ and \sum_{r=1}^{R} \lambda_r = 1$ are not required for the iterative weight calculation in (13). In order to guarantee the reliability, τ should be set relatively larger.

2.3. Dynamic Weight Interpolation (DWI)

The calculation of (12) causes no trouble for server-based ASR. However, computation and memory demands prohibit its use on low–power portable devices. In these cases, a simple weight interpolation method based on speech-conditioned prior distortion is adopted. Given the marginal density of X_t, the instantaneous distortion can be denoted as integration over speech space.

$$\begin{aligned} &\mathbf{b}_t = \int_{X_t} g(X_t)d(X_t) \quad where \\ &g(X_t) = \iint_{N_t,H_t} f(X_t,N_t,H_t)\phi(X_t,N_t,H_t)dN_t dH_t \end{aligned} \tag{13}$$

An appropriate approximation of this integration is the sum of conditional expectations of $\mathbf{b}_t$ given $\boldsymbol{\omega}_m$. If the conditional expectations of $\mathbf{b}_t$ at these given points could be denoted as:

$$\mathbf{B}_M = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_M] \ where \ \mathbf{b}_\mathrm{m} = \mathbf{B} \cdot \overline{\boldsymbol{\lambda}}_m \tag{14}$$

, the instantaneous distortion can be calculated as:

$$\mathbf{b}_t \approx \sum_{m=1}^{M} \mathbf{b}_m \cdot \gamma_{m,t} = \sum_{m=1}^{M} \mathbf{B} \cdot \overline{\boldsymbol{\lambda}}_m \cdot \gamma_{m,t} \tag{15}$$

where $\overline{\boldsymbol{\lambda}}_m$ is the conditional weight-expectation at given points $\boldsymbol{\omega}_m$. The instantaneous weights on $\mathbf{B}$ can be calculated as a linear combination of these pre-calculated weight-expectations:

$$\boldsymbol{\lambda}_t = \sum_{m=1}^{M} \gamma_{m,t} \cdot \overline{\boldsymbol{\lambda}}_m \tag{16}$$

2.4. PDC with Speech-conditioned Prior Evolution (PDC-SPE)

Due to the insufficient observations in short-time analysis, the calculation of PDC is heavily dependent on prior distribution. In the discussion of 2.2, a fixed single Gaussian is assumed for the prior weight distribution. This crude assumption prevents PDC from obtaining good performance in experiments.

In previous studies [9,10], through specifying the prior density as a more informative conjugate prior, a reproducible prior/posterior pair can be derived analytically. Different from those complex methods, a speech-conditioned prior evolution scheme is introduced to PDC. The time-variant average weights of DWI can be plugged into (12) in place of the fixed prior weight-expectation $\bar{\boldsymbol{\lambda}}$. Alternatively, an accumulated weight-interpolation is used:

$$\boldsymbol{\lambda}_t = \alpha\bar{\boldsymbol{\lambda}} + \frac{1-\alpha}{t}\sum_{m=1}^{M}\sum_{j=0}^{t}\gamma_{m,t}\cdot\bar{\boldsymbol{\lambda}}_m \quad where \quad 0\le\alpha\le 1 \qquad (17)$$

The instinctive purpose of this method is to incrementally adapt the initial weights to trace the newest scenario changes. Since the prior evolution works on incremental mode, no need to store previous consecutive data as in short-time based methods. This is equal to adjusting the prior mean vector of (11) according to different speech segments, while the prior-variance-matrix is kept unchanged. For this new method, incrementally refreshed prior statistics and the data of current window jointly contribute to the estimation of instantaneous distortion.

3. SCENARIO-ORIENTED PRIOR INFORMATION

For the Bayesian learning, appropriate prior statistics is quite crucial. For the proposed methods in the previous part, there are two kinds of prior information that should be pre-calculated in the training phase: first, the distortion pattern-set and second, the initial weights on these patterns.

The pattern-set can be derived through data-driven approach, whereas incorporating some expert scenario-classification would be helpful. The following four conditions have been chosen as representatives of mobile environments:

(i) Indoor environment (Home/office);
(ii) Public place (background noise);
(iii) Pedestrians by road side/at bus stop;
(background traffic noise)
(iv) Passengers in moving cars, railways, buses, etc.
(background traffic noise and engine noise)

It is practically impossible to consider more than the above four acoustic conditions. Any further differentiation would only deliver sparse observations of training data. There is about an hour's data in each pre-defined scenario collected and transcribed by hand. Supervised MLLR with a global bias transform is conducted on utterance level to calculate environmental distortions. The average distortions in power spectrum are depicted in figure (1).

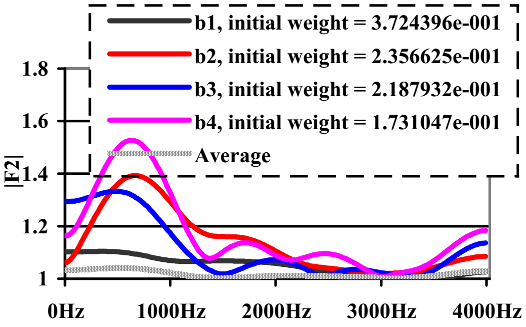


Figure (1). Distortion patterns in power spectrum

The initial weights for these patterns are accumulated by their occupational counts across all training data, **in which most utterances have no information about recording environments.**

$$\bar{\lambda}_r = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{t=1}^{T}\gamma_{m,n,r,t}}{\sum_{j=1}^{R}\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{t=1}^{T}\gamma_{m,n,j,t}} \quad 1\le r\le R \qquad (18)$$

where $\gamma_{m,n,r,t}$ is the post probability of single Gaussian $\boldsymbol{\omega}_{m,n}$ on observation $\mathbf{o}_t$ with the scenario assumption $\mathbf{b}_r$. And speech-conditioned initial weights at given points $\boldsymbol{\omega}_m$ are accumulated on corresponding observations:

$$\bar{\lambda}_{m,r} = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T}\gamma_{m,n,r,t}}{\sum_{j=1}^{R}\sum_{n=1}^{N}\sum_{t=1}^{T}\gamma_{m,n,j,t}} \quad 1\le r\le R \;\; 1\le m\le M \qquad (19)$$

In Figure (1), it should be noted that while the average distortion is neglectable (see the faded line), the distortion of each single pattern couldn't be neglected (see the solid lines). This implies that severe phase distortion exists among these distortions. Phase-frequency analysis reveals that it is caused by the so called "dispersion effect" among these patterns. Thus, the scenario-specific information is masked in acoustic model if we pool training data from different scenarios. Since it is uneconomic and pointless to train scenario-specific models for different applications, pattern-based compensation is the feasible solution.

4. EXPERIMENTS AND DISCUSSIONS

The recognizer used throughout our experiments is the InfoStar3.0, InfoTalk's multilingual ASR system especially designed for DSR applications. The 39-dimension front-end consists of 12 ACELPC (the Algebraic Code Excited Linear Prediction Coding recommended by ETSI for 3G networks, [11]) cepstral, cepstral energy and their first and second order derivatives. It is a phonetically mixture-tied system. Hundreds of hours of speech data, both "clean speech" collected indoors and "noisy speech" collected outdoors, are used for acoustic training. To compress the memory size for real applications, a within-syllable tri-phone structure is adopted for acoustic modeling. The acoustic model for each language contains more than 15,000 Gaussians.

Some prevailing methods for noisy speech recognition and different versions of PDC are investigated in this comparison:

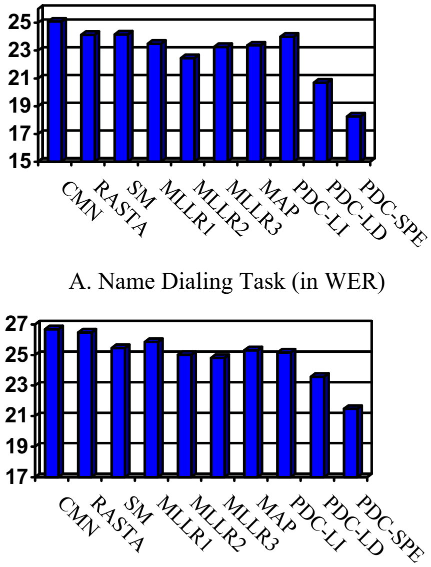
1. CMN, which serves as the baseline;
2. RASTA introduced in [2];
3. SM introduced in [4];
4. MLLR1, unsupervised mean-adaptation with a global block-diagonal matrix transformation;
5. MLLR2, unsupervised mean-adaptation with 5 block-diagonal matrices transformation;
6. MLLR3 unsupervised mean-adaptation with 10 block-diagonal matrices transformation;
7. MAP introduced in [6];
8. PDC-LI, PDC with Language-Independent initial weights. The numerator and denominator in (18) are summed across all training data regardless of language attribute;
9. PDC-LD, Language Dependent PDC. Language-specified initial weights are calculated only using the data from corresponding language.
10. PDC-SPE, PDC with Speech-conditioned Prior Evolution.

Two language-dependant tasks are investigated separately.

A. Name Dialing Task (in Mandarin)
It contains 500 Chinese Names. Name length varies between 2~4 Mandarin syllables.

B. Command & Control Task (in English)
It contains about 500 commonly used phrases and words for grammar-based command and Control applications.

Test data are collected through full rate GSM phones in outdoor environments. Each test-set contains 2,000 utterances from 50 callers (half from males, half from females). All callers are native speakers. Since some methods in comparison (MLLR and MAP) need extra data for adaptation. Five utterances from every caller are excluded from the testing-sets to serve as adaptation data.



A. Name Dialing Task (in WER)

B. Command & Control Task (in WER)
Figure (2). Performance Comparison (in WER)

Among the three MLLR methods, MLLR2 has an obvious improvement. More complex transformation shows no obvious advantage. Those unseen models in adaptation data limit the performance of MAP. PDC-LD outperforms PDC-LI obviously. This reveals that distortion has close relations with the language information (or the speech signal). It is this finding that prompted us to integrate speech-conditioned prior knowledge with PDC.

The new PDC methods are quite flexible that only R weights need to be adjusted to deal with a wild variety of scenario mismatches. Since normally R<<D, which is the feature size, the new method has a more simplified parameterization than most of other optimal estimation methods for noisy speech recognition. Unlike model adaptation, the new method works on testing data itself at run time. No extra data for adaptation is needed. It is able to trace the changing conditions during recognition. In the above two tasks, PDC-SPE provides 26% and 19.5% relative reduction in WER over corresponding baseline respectively. With respect to MLLR2, PDC-SPE achieves a relative improvement of 15% in WER.

Compared with adaptive model combination [12], the new proposed method is easier to implement since the scenario-oriented distortion patterns require fewer training data. It is economic in computation and memory considerations.

Within the scope of this study, the instantaneous distortion is calculated as a time-variant combination of the pre-calculated patterns. Since the major part of environmental effects is kept in the pattern-set and intact throughout the calculation, the online adjustment of combining weight is not enough to reflect serious variations. If the working scenario is far from those within the pattern-set, we hardly expect that the new method could perform as well as it is in this experiment. How to deal with unseen scenarios is still our future work.

5. CONCLUSIONS

Obstacles to robust speech recognition in mobile environments include acoustical degradations produced by additive noise, channel filtering, nonlinearities in coding/decoding, as well as impulsive interfering sources. While it is difficult to tackle such a wild range of degradations with limited data, it is feasible to compensate the degraded speech with some prior distortion knowledge from similar applications. In this study, the PDC method is proposed to estimate the instantaneous distortion as a time-variant linear combination of several distortion patterns, which are calculated in training phase to represent environmental characteristics of typical application scenarios. And PDC-SPE, the enhanced version of PDC with speech-conditioned initial weight evolution, shows obvious advantage over previous methods.

6. REFERENCES

[1] J.M. Huerta, R.M. Stern, "Instantaneous Distortion based Weighted Acoustic Modeling for robust speech recognition of coded speech", ICSLP 2000.

[2] H. Hermansky, N. Morgan, RASTA Processing of Speech, *IEEE Trans on SAP*, Oct, 1994.

[3] Mazin G. Rahim, and Biing-Hwang Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition", *IEEE Trans. on SAP*, Vol. 4, No. 1, pp. 19-30, Jan, 1996.

[4] A.Sankar and C.H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", *IEEE Trans. on SAP*, Vol. 4, No. 3, pp. 190-202, May 1996.

[5] C. J. Leggetter and P. C. Woodland, "Speaker adaptation of continuous density HMMs using multivariate linear regression, " in Proc. of ICSLP 94, Japan, Sep.1994.

[6] Gauvain, J.-L., Lee, C.-H., 1994. "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Trans on SAP*, Vol.2, 291-298.

[7] M. Gales. Model-Based Techniques for Noise Robust Speech Recognition. PhD thesis, Cambridge University, 1995.

[8] Huayun Zhang, Zhaobing Han, Bo Xu "Code Book Dependent Dynamic Channel Estimation For Mandarin Speech Recognition Over Telephone", ICSLP'02.

[9] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans.on SAP*, vol. 5, pp. 161–172, Mar. 1997.

[10] J.T. Chien, "Online Hierarchical Transformation of Hidden Markov Models for Speech Recognition", *IEEE Trans. On SAP*, Vol. 7, no. 6, pp. 656-667, November 1999.

[11] http://webapp.etsi.org/action/PU/20050726/ts_126190v060101p.pdf

[12] C. Huang, T. Chen and E. Chang, "Adaptive Model Combination for Dynamic Speaker Selection Training," in Proc. ICSLP2002, vol. 1, pp. 65-68, 2002.