Polynomial-Approximation-Based Model Adaptation for Noisy Speech Recognition [*]

◎ Chandra Kant Raut Hitoshi Yamamoto Takuya Nishimoto Shigeki Sagayama
(Graduate School of Information Science and Technology, The University of Tokyo)

Abstract

This paper addresses the issue of estimating HMM model parameters for noisy speech from clean HMM and noise statistics. The model parameters are estimated by approximating the non-linear function governing the relationship between speech and noise, by a Lagrange polynomial. The results showed significant improvement in performance with this approach.

1 Introduction

The performance of speech recognizers trained with clean speech degrades when used in noisy environment, due to mismatch between training and testing conditions. Parallel Model Combination (PMC) [6] has been an effective method to cope with this robustness issue. PMC estimates the model for noisy acoustical environment by combining clean HMM and noise HMM, and thus reduces the mismatch between training and testing conditions. However, an accurate estimation of model parameters involves numerical integration, which is computationally very expensive. Data-driven PMC [5], that is based on generating samples of corrupted speech vectors by Monte-Carlo simulation, is sufficiently accurate compared to numerical integration, but still slow. Other approximations such as log-normal, log-add and log-max are computationally efficient, but less accurate [4]. PMC log-normal approximation, that is most commonly used, assumes that the sum of two log-normally distributed random variables is itself log-normally distributed [6].

Vector Taylor Series (VTS) [1, 2] is yet another approach to combine the models by approximating the non-linear relationship between speech and noise with a truncated vector Taylor series. However, other polynomials optimized to approximate the parameters of distribution can give better result than the Taylor series [2].

In this paper, we approximate the non-linear function governing the relationship between speech and noise by a Lagrange polynomial, and then estimate the model parameters for noisy speech. The accuracy of the approximation is compared with other methods, and the performance of speech recognizers with Lagrange polynomial approximation is also evaluated.

2 Model of the Environment

An acoustical model demonstrating the effect of additive noise $n[m]$ and channel filtering $h[m]$ over a clean speech signal $x[m]$ is shown in Figure 1.

The corrupted speech is given by:

$$y[m] = x[m] * h[m] + n[m] \tag{1}$$

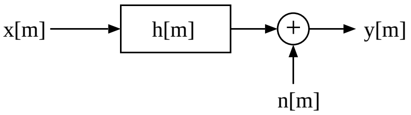

Figure 1: Model of the acoustical environment.

where m is sample number. In power spectral domain:

$$|Y(f)|^2 \approx |X(f)|^2|H(f)|^2 + |N(f)|^2 \tag{2}$$

$$\Rightarrow \ln|Y(f)|^2 \approx \ln|X(f)|^2 + \ln|H(f)|^2 + \ln\left(1 + e^{\ln|N(f)|^2 - \ln|X(f)|^2 - \ln|H(f)|^2}\right) \tag{3}$$

$$\Rightarrow y = x + h + \ln(1 + e^{n-x-h}) \tag{4}$$

where x, n, h and y represent log-spectral energies of clean signal, additive noise, convolutive noise and corrupted signal respectively, for given frequency f.

Thus, the relationship between speech and noise is non-linear one, as given in Eq.(4). Experiments show that even if noise and clean speech parameters have Gaussian distribution (in log-domain), the corrupted speech parameters do not have Gaussian distribution anymore. However, if parameters have low variances, and in case a number of mixtures of Gaussians are used to model their distributions, the distribution of parameters can be still assumed to be Gaussian without much loss of accuracy and being able to use the same decoder optimized for Gaussian distribution.

3 Polynomial Approximation

The goal is to find the distribution (mean and variance) of noisy speech parameter y, given the distributions for noise parameters n and h, and clean speech parameter x. First, mean, i.e. expectation value, of y is given as:

$$\begin{aligned} E(y) &= E(x) + E(h) + E[ln(1 + e^{n-x-h})] \\ &= E(x) + E(h) + E[g(x, n, h)] \end{aligned} \tag{5}$$

Provided x, n and h have Gaussian distributions, $E[g(x, n, h)]$ does not have any closed-form expressions.

Now, the task here is to approximate the function $g(x, n, h)$ by some polynomial that can closely approximate it within a given range, with its low order as far as possible. The polynomial approximation can be simplified by reducing function $g(x, n, h)$ to univariate one, by taking $z = n - x - h$, where $z \sim \mathcal{N}(\mu_n - \mu_x - \mu_h, \sigma_n^2 + \sigma_x^2 + \sigma_h^2)$.

We choose here 2nd-order Lagrange polynomial to approximate the function $g(z)$, as given by:

$$g(z) \approx P_2(z) = \sum_{k=0}^{2} g(z_k) \prod_{i=0, i\neq k}^{2} \frac{(z - z_i)}{(z_k - z_i)} \tag{6}$$

[*] 多項式近似に基づく雑音下音声認識のためのモデル適応, チャンドラ カント ラウト, 山本 仁, 西本 卓也, 嵯峨山 茂樹 (東大・情報理工)

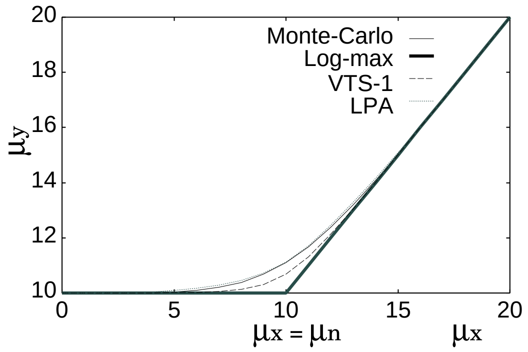

Figure 2: Estimated mean of corrupted speech: by Monte-Carlo simulation, Log-max approximation, Vector Taylor Series-1, and Lagrange Polynomial Approximation (LPA) for $\mu_n = 10$, $\sigma_n^2 = 0.1$, $\sigma_x^2 = 6$ and μ_x varying from 0 to 20.

The points z_0, z_1 and z_2 can be specified manually (one point at $z = \mu_z$ and other two chosen to minimize error in the required range), or instead, Chebyshev-Lagrange polynomial can be used that specifies the points itself.

Finally, Eq.(6) reduces to $g(z) = az^2 + bz + c$ form, where a, b and c are constants. Therefore:

$$E[g(z)] = a(\sigma_z^2 + \mu_z^2) + b\mu_z + c \qquad (7)$$

The accurate value of mean is more important than that of variance. Therefore, the covariance matrix can be retained as it is for the clean speech. However, expression for adapting diagonal variance can be derived from above approximation, in terms of higher-order moments (up to 4th moment) of z, and diagonal variances can be adapted as well.

4 Analysis of the Approximation

To analyze the accuracy of the approximation, and to compare it with other methods, a set of one-dimensional vectors was generated for speech signal by Monte-Carlo simulation. These speech vectors were corrupted by adding noise at different SNR. Noise vectors were also generated by Monte-Carlo simulation.

The mean of corrupted speech estimated by different methods has been plotted in Figure 2. The result shows that Lagrange polynomial approximation gives almost the same result as given by Monte-Carlo simulation, and thus outperforms VTS-1 and log-max approximations.

5 Experimental Results

For evaluation of Lagrange polynomial approximation based approach, the system was tested on an isolated word recognition task trained with 2620 words of same speaker taken from A-Set of ATR Speech Database. The test set consisted of 655 words of the same speaker taken exclusively from the database.

The baseline system comprised of 41 context-independent continuous-density phone HMMs with single-mixture 123 states and 26 dimensional vector composed of static and delta MFCC with energy and delta-energy coefficients. Julian 3.4 was used as the decoder. The baseline word recognition accuracy for clean speech was 93.8%.

Exhibition noise from JEITA database was added to test data at 0 dB, 5 dB, 10 dB, 20 dB and 40 dB SNRs. The word recognition accuracy reduced to 2.8% at 0 dB SNR, when recognized with clean HMMs.

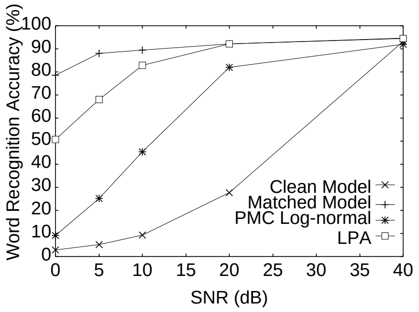

Figure 3: Recognition result with clean model, matched model, and the models adapted by PMC Log-normal and Lagrange Polynomial Approximation (LPA).

Recognition were then performed with the models adapted by Lagrange polynomial approximation. The models were adapted only for 12 static mean parameters.

Figure 3 shows word recognition accuracies at different SNRs obtained with various models. The matched models were built by training HMMs from training data corrupted by noise at given SNRs. In case of PMC log-normal approximation, both means and variances of 12 static parameters were adapted. As seen in the figure, the performance obtained with Lagrange polynomial approximation based model adaptation is close to that obtained with matched models, at high SNR; and is significantly improved compared to PMC log-normal approximation at low SNR.

6 Conclusion

The model parameters for corrupted speech can be accurately estimated by approximating the nonlinear function by a Lagrange polynomial as described in this paper. The performance of speech recognizer with Lagrange polynomial based model adaptation was significantly improved compared to other methods.

Future work includes estimation of complete covariance matrix, and evaluation of the approach on different tasks and with different models, such as with context-dependent phone models and Gaussian mixtures models. Furthermore, possibilities of using other polynomial expansions will be investigated.

References

[1] A. Acero *et al.*, "HMM Adaptation using Vector Taylor Series for Noisy Speech Recognition," Proc. ICSLP-2000, Vol. 3, pp. 869-873, 2000.

[2] P. J. Moreno, B. Raj and R. M. Stern, "A Vector Taylor Series Approach for Environment Independent Speech Recognition," Proc. ICASSP-1996, pp. 733-736, 1996.

[3] P. J. Moreno, "Speech Recognition in Noisy Environments," Ph.D. Thesis, Carnegie Mellon University, 1996.

[4] Y. Gong, "A Comparative Study of Approximations for Parallel Model Combination of Static and Dynamic Parameters," ICSLP-2002, pp. 1029-1032, 2002.

[5] M. J. F. Gales and S. J. Young, "A Fast and Flexible Implementation of Parallel Model Combination," Proc. ICASSP-1995, pp. 133-136, 1995.

[6] M. J. F. Gales, "Model-Based Techniques for Noise Robust Speech Recognition," Ph.D. Thesis, Cambridge University, 1995.