Model Composition by Lagrange Polynomial Approximation for Robust Speech Recognition in Noisy Environment

Chandra Kant Raut, Takuya Nishimoto, Shigeki Sagayama

Graduate School of Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 JAPAN
{raut, nishi, sagayama}@hil.t.u-tokyo.ac.jp

Abstract

This paper presents a technique for estimating HMM model parameters for noisy speech from given clean speech HMM and noise HMM. The model parameters are estimated by approximating the non-linear function governing the relationship between speech and noise, by a Lagrange polynomial, and thus enabling the distribution of corrupted speech parameters to have a closed form. The method is computationally efficient, and the experimental results showed significant improvement in recognition performance of noisy speech with this approach. Typically, word accuracy increased from 9.2% with clean model to 82.8% with the model composed by the proposed method as compared to 45.4% with the model composed by PMC Log-normal approximation, on an isolated word recognition task for exhibition hall noise added at 10 dB SNR.

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. The difference between training and testing conditions may result from background noise, channel distortion, speaker's stress and other factors as well. A number of techniques have been developed to deal with this robustness issue, that can be broadly categorized into robust front-ends, multiple microphones, enhancement techniques and model-based compensation schemes. This paper deals with the model-based or HMM composition approach as considered in works [2] by Varga et al., [3] by Martin et al. and [4] by Gales and Young. Model composition method estimates the model for noisy acoustical environment by combining clean HMM and noise HMM, and thus reduces the mismatch between training and testing conditions.

Parallel Model Combination (PMC) [4, 5] has been an effective method to cope with the robustness issue, and has been extensively studied. Many variations of PMC exist that attempt to estimate noise-adapted model from clean speech HMM and noise HMM. However, an accurate estimation of model parameters by PMC involves numerical integration, which is computationally very expensive. Data-driven PMC [6], 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 [7]. Further, 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 [5].

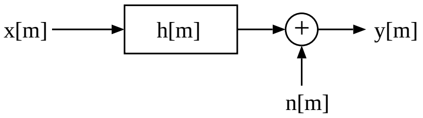


Figure 1: *Model of the acoustical environment.*

Jacobian approach to model adaptation, proposed by Sagayama *et al.* in [8], attempts to compensate model by Jacobian matrices with the difference between assumed and observed noise cepstra. However efficient and effective, the method requires some training data, and assumes that cepstral difference and variance of mixtures stay within the linearity range.

Use of neural network (NN) for combining clean speech HMM and noise HMM has been investigated in [9]. Neural networks are used to learn the nonlinearity involved in combination, and thus to produce noise-adapted HMM, by using clean speech HMM, noise HMM and SNR as inputs. Neural networks need to be trained first, using a set of input and output HMMs. Output HMM for training is obtained by a combination of MLLR, MAP and VFS adaptation techniques for a particular combination of inputs, viz. clean speech HMM, noise HMM and SNR[9]. The method has been found to be effective, however it involves building large number of sample output noisy HMMs and training of NNs, that is slow, computationally inefficient and a tedious task.

Vector Taylor Series (VTS) [10, 11] 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 [11].

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 the Lagrange Polynomial Approximation method is also evaluated on a speaker-dependent isolated word recognition task.

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:

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

where m is sample number. In power spectral domain, the filter-bank energies is given as:

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

$$\Rightarrow \ln|Y(f)|^2 \approx \ln|X(f)|^2 + \ln|H(f)|^2$$

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

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

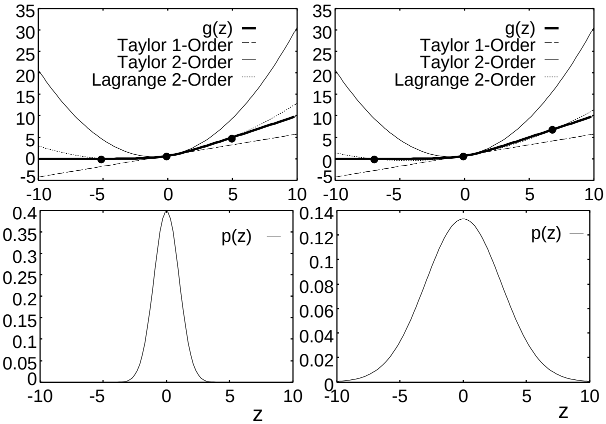
where x, n, h and y represent log-spectral energies of clean signal, additive noise, convolutive noise and corrupted signal respectively.

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 the same decoder optimized for Gaussian distribution can be used.

3. Polynomial-approximation

The goal under consideration 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:

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



(a) (b)

Figure 2: *Comparison of Polynomial Approximations*

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

Therefore, to find the value of $E[g(x,n,h)]$, we first expand the function $g(x,n,h)$ into a 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 $z \sim \mathcal{N}(\mu_n - \mu_x - \mu_h, \sigma_n^2 + \sigma_x^2 + \sigma_h^2)$.

In this work, 2nd-order Lagrange interpolating polynomial is used to approximate the function $g(z)$, given as:

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

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.

Figure 2 shows different polynomials used to approximate function $g(z) = \ln(1 + e^z)$ when $\mu_z = 0$. In Figure 2a, the points selected for Lagrange polynomial expansion are $z_0 = \mu_z$, $z_1 = z_0 - 5$ and $z_2 = z_0 + 5$. As seen in the figure, Lagrange polynomial has been able to approximate the function up to larger range and more accurately than even 2-nd order Taylor's series. When variance of z is low, we can keep points z_1 and z_2 closer to z_0; however, when z has large variance, the points should be extended farther from z_0. However, extending them too farther introduces inaccuracies in the approximation in region close to $z = z_0 = \mu_z$, where most of data occurs. Therefore, the points z_1 and z_2 should be placed at some optimum values depending on the variance of z.

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

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

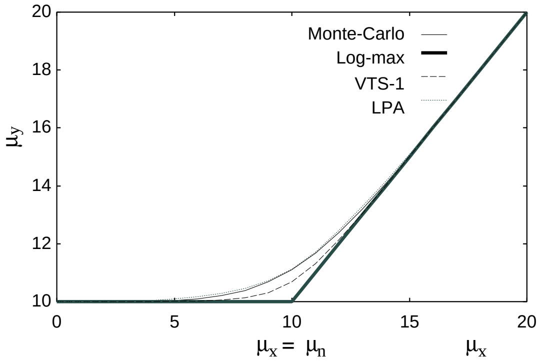


Figure 3: *Estimated mean of corrupted speech: by Monte-Carlo simulation, Log-max approximation, Vector Taylor Series-1, and Lagrange Polynomial Approximation (LPA) for $\mu_n = 10$, $\sigma_n^2 = 0.1$, $\sigma_x^2 = 6$ and μ_x varying from 0 to 20.*

Using this estimated value of $E[g(z)]$ in Eq. 5, the mean for corrupted speech vector can be computed. As, the accurate value of mean is more important than that of variance, the covariance matrix can be retained as it is for the clean speech. However, expression for adapting diagonal variances 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.

The method for estimating model parameters for corrupted speech is shown in Figure 4. As approximation is done in log-spectral domain, the HMM parameters of clean speech and noises in cepstral domain are converted into log-spectral domain by taking inverse DCT. This conversion of parameter vector from cepstral to log-spectral domain requires knowledge of C_0. In case if the given model parameters do not include C_0, it can be computed, as worked out in [12], by noticing the fact that the sum of the energies of Mel bands in linear spectral domain equals to total frame energy.

The statistics to account for channel distortions can be obtained by using EM based approach by maximizing the likelihood score as described in [13]. Some adaptation data are required to estimate the statistics for channel distortion.

In all cases, only diagonal elements of covariance matrix of speech and noise HMMs are considered, in order to avoid the complexity and reduce the computational expense of the algorithm.

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 parameter by Monte-Carlo simulation. These speech vectors were corrupted by adding noise at different SNR. Noise vectors were also generated by Monte-Carlo simulation.

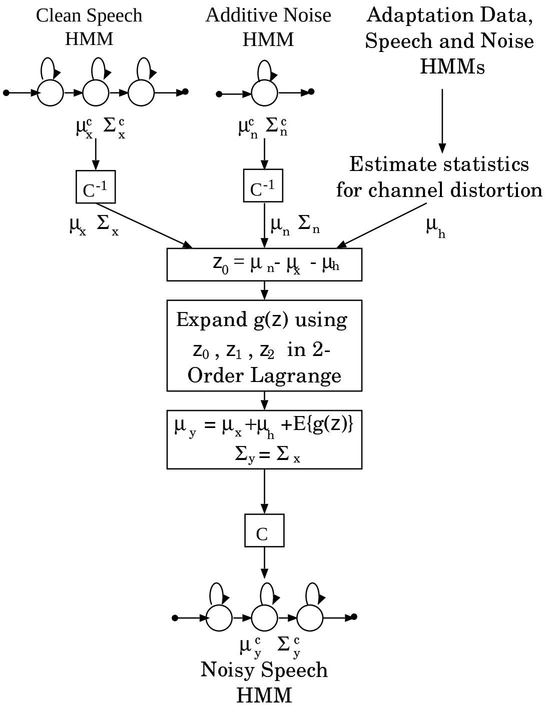


Figure 4: *Model Composition by Lagrange Polynomial Approximation*

The means of corrupted speech estimated by different methods have been plotted in Figure 3. The result shows that the Lagrange polynomial approximation (LPA) outperforms VTS-1 and Log-max approximations. The mean estimate given by Lagrange polynomial approximation is almost same as given by Monte-Carlo simulation, however cutting down the computational cost to large extent.

5. Experimental results

To evaluate Lagrange polynomial approximation based approach, it 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 from the same speaker taken exclusively from the database.

The baseline system comprised of total 41 context-independent continuous-density single-mixture phone HMMs with 123 states in total, and 26-dimensional vectors composed of 13-MFCCs (with C_0) and their deltas were used as input features. The baseline word recognition accuracy for clean speech was 93.8%, with Julian 3.4 used as decoder.

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

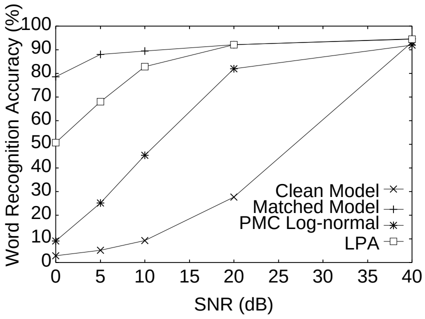


Figure 5: *Recognition result with clean model, matched model, and the models adapted by PMC Log-normal and Lagrange Polynomial Approximation (LPA).*

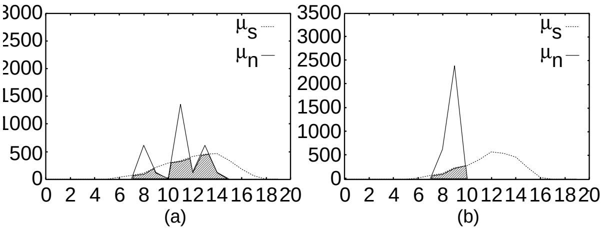


Figure 6: *Histogram of Speech and Noise Parameters' means in log-spectral domain*

SNR, when recognized with clean HMMs.

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

Figure 5 shows word 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 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.

Figure 3 shows that when $\mu_n >> \mu_x$ or $\mu_n << \mu_x$, any of the methods can estimate μ_y with sufficient accuracy. However, if $\mu_n \approx \mu_x$, other methods fail to give accurate estimate, whereas LPA works very well. Therefore, when HMM parameters of noise and speech fall close to each-other during combination, LPA's advantage will be more pronounced. Figure 6 shows the histogram of parameters' means (in log-spectral domain) of speech and noise. In the case shown in Figure 6a, speech means and noise means occur closer to each-other(shaded area) than in case of Figure 6b. Thus improvement obtained with LPA will be more noticeable with HMMs of the case (a) than in case (b), compared to other methods.

6. Conclusion

The model parameters for corrupted speech can be accurately and efficiently estimated by approximating the non-linear function by a Lagrange polynomial as described in this paper. The performance of speech recognizer with Lagrange polynomial approximation based model composition was significantly improved compared to other methods.

Future work includes estimation of 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.

7. References

[1] Raut, C. K., Yamamoto, H., Nishimoto, T., and Sagayama, S., "Polynomial-Approximation-Based Model Adaptation for Noisy Speech Recognition," in *Proc. The 2004 Spring Meeting of The Acoustical Society of Japan*, vol. 1, pp. 121-122, 2004.

[2] Varga, A. P. and Moore, R. K., "Hidden Markov Model Decomposition of Speech and Noise," in *Proc. ICASSP90*, pp. 845-848, 1990.

[3] Martin, F., Shikano, K., and Minami, Y., "Recognition of Noisy Speech by Composition of Hidden Markov Models," in *Proc. Eurospeech93*, pp. 1031-1034, 1993.

[4] Gales, M. J. F. and Young, S. J., "Robust Continuous Speech Recognition using Parallel Model Combination," in *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, pp. 352-359, 1996.

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

[6] Gales, M. J. F. and Young, S. J., "A Fast and Flexible Implementation of Parallel Model Combination," in *Proc. ICASSP95*, pp. 133-136, 1995.

[7] Gong, Y., "A Comparative Study of Approximations for Parallel Model Combination of Static and Dynamic Parameters," in *Proc. ICSLP02*, pp. 1029-1032, 2002.

[8] Sagayama, S., Yamaguchi, Y., Takahashi, S., and Takahashi, J., "Jacobian Approach to Fast Acoustic Model Adaptation," in *Proc. ICASSP97*, vol. 2, pp. 835-838, 1997.

[9] Furui, S. and Itoh, D., "Neural-Network-Based HMM Adaptation For Noisy Speech," in *Proc. ICASSP01*, vol.1, pp.365-368, 2001.

[10] Acero, A. *et al.*, "HMM Adaptation using Vector Taylor Series for Noisy Speech Recognition," in *Proc. ICSLP00*, vol. 3, pp. 869-873, 2000.

[11] Moreno, P. J., Raj, B., and Stern, R. M., "A Vector Taylor Series Approach for Environment Independent Speech Recognition," in *Proc. ICASSP96*, pp. 733-736, 1996.

[12] Crafa, S., Fissore, L., and Vair, C., "Data-Driven PMC and Bayesian Learning Integration for Fast Model Adaptation in Noisy Environment," in *Proc. ICSLP98*, vol. 2, pp. 471-474 , 1998.

[13] Minami, Y. and Furui, S., "A Maximum Likelihood Procedure for a Universal Adaptation Method Based on HMM Composition," in *Proc. ICASSP95*, vol. 1, pp. 129-132, 1995.