AN ACCURATE NOISE COMPENSATION ALGORITHM IN THE LOG-SPECTRAL DOMAIN FOR ROBUST SPEECH RECOGNITION

Mohamed Afify

Department of Information Technology
Faculty of Information and Computer, Cairo University

ABSTRACT

This paper presents an algorithm for noise compensation in the log-spectral domain. The idea is based on the use of accurate approximations which allow theoretical derivations of the noisy speech statistics, and using these statistics to define a compensation algorithm under a Gaussian mixture model assumption. The algorithm is tested on a digit data base recorded in the car, the word recognition accuracies for the baseline (uncompensated), first order VTS, the proposed method, and the matched test, are 85.8%, 90.6%, 93.1%, and 93.9% respectively. This clearly indicates the performance gain due to the proposed technique.

1. INTRODUCTION

The input of a speech recognition system is often corrupted by additive noise. This usually leads to severe performance degradation especially when the baseline system is trained under clean conditions. There has been considerable research to reduce this deterioration, and hence increase the robustness of speech recognition systems to additive noise.

Many techniques approach the problem in the log spectral domain, a closely related space to the widely used Mel frequency cepstral coefficients (MFCC). The basic idea is to form a mismatch function which represents the noisy speech in the log spectral domain. It is known that additive noise in the linear spectral domain leads to a non linear mismatch function in the log spectral domain, and hence the derivation of the exact noisy speech statistics becomes difficult, and some approximations must be used. One of the most popular approaches in this context is the vector Taylor series (VTS). In the vector Taylor series (VTS) approach [3], a series expansion is used to approximate the nonlinear mismatch function, and thus facilitates both the calculation of the statistics for the noisy speech and applying minimum mean square error (MMSE) estimation to retreive the clean signal.

In this paper we first formulate noise compensation in the log spectral domain. We then point out that if the clean speech is represented by a Gaussian mixture model, and if the Gaussian mixture assumption is still valid after noise corruption, a minimum mean square error compensation algorithm is completely defined by three statistics for each Gaussian component. These statistics are, the mean and variance of noisy speech, and the covariance of the clean and noisy speech. Using the mismatch function of [1] we derive expressions for these statistics, under some approximations to be discussed in the paper, and hence develop a new compensation algorithm in the log spectral domain. In contrast to universal approximations like VTS, the approximations employed are tailored to the mismatch function, and hence are expected to lead to low approximation error. In addition, they allow deriving mathematical expressions of the required statistics, as will be discussed in the paper.

The paper is organized as follows. Section 2 formulates parametric noise compensation in the log spectral domain. The main results of the paper are given in section 3. We derive expressions for the mean and variance of noisy speech, and the covariance of clean and noisy speech, for the mismatch function of [1] under some approximations. Section 4 gives some implementation issues related to the compensation algorithm and comments on the computational complexity of the proposed method. Experimental results are given in Section 5. Finally, conclusions are given in Section 6.

2. PARAMETRIC NOISE COMPENSATION IN THE LOG-SPECTRAL DOMAIN

This section formulates the problem of parametric noise compensation in the log-spectral domain which is the basis for the proposed noise compensation algorithm, and introduces vector Taylor series (VTS) which is a very popular approximation used in this context.

As we assume components are independent the presentation is in scalar form, and the extension to the vector case is by simply repeating the same argument for every vector dimension. Assume that speech is corrupted by additive noise in the time(linear-spectral) domain. Denote the noisy speech, clean speech, and noise in the log-spectral domain by y, x, and n respectively. The mismatch function in the log-spectral domain can be written as [1]

$$y = x + \log(1 + \exp(n - x)). \tag{1}$$

We start by assuming parametric models for the speech and noise in the log-spectral domain. A very popular choice is to use a Gaussian mixture of size M with mixture weights, means and variances $\{c_m, \mu_{xm}, \sigma^2_{xm} \mid 1 \leq m \leq M\}$ for the clean speech, and a Gaussian with mean μ_n and variance σ^2_n for the noise. By making the assumption that under the transformation in Equation (1) the noisy speech still follows a Gaussian mixture of size M having new parameters $\{c_m, \mu_{ym}, \sigma^2_{ym} \mid 1 \leq m \leq M\}$[1] , the compensation problem reduces to the following two tasks:

- Use the noisy speech pdf to get an estimate of the

[1]This is a reasonable assumption as a Gaussian mixture of sufficient size can approximate any pdf.

noise mean (starting from an initial estimate)

$$\hat{\mu_n} = \arg\max_{\mu_n} p(y) = \arg\max_{\mu_n} \sum_{m=1}^{M} c_m \mathcal{N}(y, \mu_{ym}, \sigma^2_{ym}) \quad (2)$$

- Obtain a minimum mean square error (MMSE) estimate of the clean speech given the noisy speech. Under the Gaussian assumption this can be calculated as

$$\begin{aligned} \hat{x} &= E[x|y] = \sum_{m=1}^{M} p(m|y) E[x|y,m] \\ E[x|y,m] &= \mu_{xm} + \sigma_{xym}/\sigma^2_{ym}(y - \mu_{ym}) \end{aligned} \quad (3)$$

where $p(m|y)$ can be easily claculated using Bayes rule.

From the above discussion, and dropping dependence on mixture component for convenience, we find that the compensation algorithm can be defined in terms of three quantities for each Gaussian component: the noisy speech mean and varaince μ_y, and σ^2_y, and the covariance of the clean and noisy speech σ_{xy}. Due to the nonlinearity of Equation (1) it is difficult to directly calculate these statistics.

A very popular approximation is to expand the function in Equation (1) using Taylor series to obtain a polynomial which facilitates the above computations [3]. The use of first order polynomial, known as first order VTS, is of particular interest in practice. In the next section we present two new approximations that allow calculating expressions of the above statistics, and hence obtaining a new noise compensation algorithm.

3. MAIN RESULTS

In this section we give the main results obtained in this paper. Derivations are omitted for space limitation and will be presented elsewhere. In particular we derive expressions, under some approximations, for the mean and variance of random variable y, and the covariance of random variables x, and y, where y is related to x, and n as follows

$$\begin{aligned} y &= x + \log(1 + \exp(n - x)) \\ &= x + \log(1 + \exp(w)) = x + f(w) \end{aligned} \quad (4)$$

The results are based on two main approximations **A1**, and **A2**, these approximations will be stated below, and will be discussed in Section 3.1.

Approximation A1

$$\begin{aligned} f(w) &= \log(1 + \exp(w)) \\ &\approx \log(1 + \exp(w)) u(w + a) \end{aligned} \quad (5)$$

where $u(w+a) = 1$ for $w \geq -a$, and $u(w+a) = 0$ otherwise, and a is a constant as will be discussed below.

Approximation A2

- $F(w) \equiv df(w)/dw = 1/(1 + \exp(-w)) \approx \Phi(w)$.
- The previous item implies that $d^2 f(w)/dw^2 \approx \phi(w)$.

where

$$\phi(w) = \frac{1}{\sqrt{2\pi}} \exp(-w^2/2) \quad (6)$$

$$\Phi(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} \exp(-m^2/2) dm \quad (7)$$

After giving the basic relationship in Equation (4), and the stated approximations we have the following theorem.

Theorem 1

Assume that x is normally distributed with mean μ_x, and variance σ^2_x, and that n is normally distributed with mean μ_n, and variance σ^2_n, and define

$$\begin{aligned} \mu &= \mu_n - \mu_x \\ \sigma^2 &= \sigma^2_n + \sigma^2_x \end{aligned} \quad (8)$$

Further assume that x, and n are statistically independent, and that y is related to x, and n according to Equation (4). We have the following results for the mean of y, μ_y, the variance of y, σ^2_y, and the covariance of x, and y, σ_{xy}:

1. Under the above conditions, and approximations $A1$, and $A2$, and for arbitrary a, where $|a| < \infty$, μ_y is given by

$$\mu_y \approx \mu_x + \frac{\sigma\phi(a)}{1 - \Phi(a)} \int_0^1 \Phi\left(\frac{\mu - \rho a \sigma}{\sqrt{1 + \sigma^2(1 - \rho^2)}}\right) d\rho \quad (9)$$

2. Under the above conditions and approximation $A2$, σ_{xy} is given by

$$\sigma_{xy} \approx \sigma^2_x \left(1 - \Phi\left(\frac{\mu}{\sqrt{1 + \sigma^2}}\right)\right) \quad (10)$$

3. Under the above conditions and approximation $A2$, σ^2_y is given by

$$\begin{aligned} \sigma^2_y \approx\ & \sigma^2_x \left(1 - \Phi\left(\frac{\mu}{\sqrt{1 + \sigma^2}}\right)\right)^2 + \sigma^2_n \Phi^2\left(\frac{\mu}{\sqrt{1 + \sigma^2}}\right) + \\ & \int_0^1 \int_0^\alpha W(\rho) d\rho d\alpha \end{aligned} \quad (11)$$

where

$$\begin{aligned} W(\rho) =\ & \frac{\sigma^4 C(\rho)}{\sqrt{1 + \sigma^2 + \sigma^2\rho^2 C^2(\rho)}} \phi\left(\frac{\mu C(\rho)(1 - \rho)}{\sqrt{1 + \rho^2 C^2(\rho)}}\right) \times \\ & \phi\left(\frac{\mu(1 + \rho C^2(\rho))}{\sqrt{1 + \rho^2 C^2(\rho)}\sqrt{1 + \sigma^2 + \sigma^2\rho^2 C^2(\rho)}}\right) \\ C(\rho) =\ & \frac{1}{\sqrt{1 + \sigma^2(1 - \rho^2)}} \end{aligned} \quad (12)$$

After presenting the results in the above theorem some comments on the used approximations will be given in Section 3.1.

3.1. Discussion of the Approximations

The proof of results in theorem 1 is based on the use of Approximations $A1$, and $A2$, and hence their accuracy are directly related to the accuracy of these approximations. Hence we will discuss the accuracy of these approximations.

- Approximation $A1$ simply multiplies $f(w)$ by a unit step at $w = -a$. Noting that $f(w)$ is a decreasing function of w which approaches zero for reasonably large negative values of w (refer to [4] for a sketch of $f(w)$). Thus, for a large enough the multiplication by the step will closely approximate the original function. Of course we need a to be as large

as possible to obtain a better approximation. However, as can be seen from Equation (9) the mean becomes undefined (this is due to technical reasons in the derivation which is omitted here) when $a \to \infty$. Thus, there is a tradeoff in choosing the value of a between the accuracy of the approximation and the computability of the expected value. In this paper we use the following method for computing a which was found to work very well in practice.

$$a = \min((3+\mu)/\sigma, 3) \tag{13}$$

This is based on the following reasoning. Defining $w = \sigma v + \mu$, where μ, and σ are the mean and standard deviation of w as defined in Equation (8), and consider that f(w) vanishes when its argument approaches -3 [2], this leads to the first argument of the min, the other argument is used to ensure that a does not take very large values, and hence guarantees the computability of the expected value.

- Approximation $A2$ simply replaces a sigmoid function with a Gaussian CDF. Both functions are in fact sigmoids with very similar behaviour (refer to [2] for a sketch which highlights this similarity). However, the use of a Gaussian CDF facilitates the calculation of some expected values. This approximation has been used in our previous work, in a similar context, to develop upper and lower bounds on the mean of noisy speech [2].

The choice of the above approximations is motivated by the properties of the target function and its derivatives as discussed in the above two items to ensure an intuitively very accurate approximation. These approximations can be considered as tailored approximations to the function in contrast to other universal approximators as vector Taylor series. In the experiments we compare the proposed approximations to the very popular first order vector Taylor series.

4. IMPLEMENTATION OF THE COMPENSATION ALGORITHM

The proposed method leads to integral expressions for calculating the noisy speech mean and variance. In this section, we will discuss how these integrals are evaluated, and hence the noisy speech statistics are calculated. We will then outline the whole compensation algorithm, and discuss its computational complexity.

4.1. Evaluation of the Integrals

Expressions for the noisy speech mean and variance, as discussed above, result in some integrals that should be numerically evaluated. There are many ways to numerically evaluate an integral. In this work we use a very simple method based on text book definition of an integral, that we found to work well in practice, and that results in the following expressions

$$\mu_y = \mu_x + \frac{\sigma\phi(a)}{N_\rho(1-\Phi(a))}\sum_{i=0}^{N_\rho-1}\Phi\left(\frac{\mu - i\Delta_\rho a\sigma}{\sqrt{1+\sigma^2(1-i^2\Delta_\rho^2)}}\right) \tag{14}$$

[2]This is a practically reasonable value to ensure good approximation.

$$\sigma_y^2 = \sigma_x^2\left(1-\Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right)\right)^2 + \sigma_n^2\Phi^2\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) + \frac{1}{N_\rho}\sum_{i=0}^{N_\rho-1}\frac{1}{i+1}\sum_{j=0}^{i}W(\Delta_\rho j) \tag{15}$$

where N_ρ is the number of bins used to evaluate the integral, $\Delta_\rho = 1/(N_\rho - 1)$, and $W()$ is defined in Equation (12). In this work we found that $N_\rho = 11$ works well in practice.

4.2. Summary of the Compensation Algorithm

In this section we summarize the whole compensation algorithm. We note that the algorithm is implemented in the log-spectral domain, and that different channels are considered independent, and hence we use scalar notation and repeat the same arguments for every vector dimension.

1. Initialize the mean μ_n, and the variance σ_n^2 of the noise pdf $p(n)$ using the first N frames of an utterance. Typically we use $N = 10$ (this corresponds to 100 msec for the used frame rate).
2. Refine the noise mean estimate using first order VTS if desired. Comments on this step will be given below.
3. For each mixture component in the clean speech Gaussian mixture model (of size M) do
   - If the models are trained in the MFCC domain map them to the log-spectral domain as in [5].
   - Calculate μ, and σ, as given in Equation (8), and a as given in Equation (13).
   - Calculate the noisy speech statistics, μ_y as given in Equation (14), σ_y^2 as given in Equation (15), and σ_{xy} as given in Equation (10).
4. For each frame in the test utterance do
   - Map the frame, initially in the MFCC domain, to the log-spectral domain.
   - Calculate the posteriors using the noisy speech Gausian mixture model and Bayes rule.
   - Estimate the clean speech as in Equation (3).
   - Map the estimated clean speech to the MFCC domain, and calculate, if desired, the delta, and delta-delta coefficients.

The initial noise estimate is usually refined using the estimated noisy speech model as in Equation (2). However, using the derived expressions of the noisy speech mean and variance will not result in a closed form solution. For this reason we use a first order VTS approximation for refining the noise mean if desired.

4.3. Computational Complexity

In this section we discuss the computational complexity of the proposed algorithm and compare it to both first order VTS, and the Monte Carlo method. Due to space limitation we only summarize the final results, and quantitative analysis will be given elsewhere. We first point out that there are two costs associated to any compensation algorithm, the cost of calculating the statistics, and that of

estimating the clean speech. The cost of calculating the statistics of the proposed method is much larger than that of first order VTS due to the numerical evaluation of the integrals. However, if the noise model is kept unchanged for sufficiently large number of frames, the total cost of both algorithms will be dominated by the cost of estimating the clean speech. On the other hand, for the Monte Carlo method the cost of calculating the statistics is much larger than the proposed method, due to the large number of iterations needed for the estimates to converge, and even larger than the cost of clean speech estimation.

5. EXPERIMENTAL RESULTS

The proposed algorithm is evaluated on a hands-free database (CARVUI database) recorded inside a moving car using a microphone array of 16 channels, and a close-talking microphone. The data was collected in Bell Labs area, under various driving conditions and noise environments. A total of 56 speakers participated in the data collection. Evaluation is limited to the digit part of the data base. The speech material from 50 speakers is used for training, and the data from the 6 remaining speakers is used for test, leading to a total of about 6500 utterances available for training and 800 utterances (comprising about 3000 digit) for test. The data is recorded at 24kHz sampling rate and is down-sampled to 8kHz and followed by a MFCC feature extraction step for our speech recognition experiments. The feature vector consists of 39 dimensions, 13 cepstral coefficients and their first and second derivatives. Cepstral mean normalization is applied on the utterance level. The recognition task consists of simple loop grammar for the digits. In our experiments, data from 2 channels only are used. The first one is the close-talking microphone (CT), the second one is a single channel from the microphone array, referred to as Hands-Free data (HF) henceforward. The average SNR is about 21dB for the CT channel and 8dB for the HF channel.

10 digit models and a silence model are built. Each model is left to right having six states, and each state has 8 Gaussian distributions. A clean speech Gaussian mixture model of size 64 is also built from the data to be used in the noise compensation. Training and testing are done using the hidden Markov model toolkit (HTK). Table 1 shows the accuracy when testing both the CT and HF data using model trained on CT data. It is clear that the recognition results fall sharply when moving to the noisy data. We then tested using the compensation algorithm and comparing it to first order VTS. We considered four scenarios: the proposed method with and without noise mean reestimation, and first order VTS with and without noise mean reestimation. In addition we performed a matched test, where a model trained on the HF training data is used. Table 2 shows the corresponding results, with the HF model labelled as Matched. By referring to Table 2 we observe that the proposed method significantly outperforms first order VTS, and is close to the matched case. Noise reestimation offers a small gain in all cases.

Test Data	Accuracy
CT	96.3
HF	85.8

Table 1: Recognition results (in %) of the close-talking (CT) microphone data and Hands-Free (HF) data using CT model.

Compensation Method	Accuracy
VTS–No Noise Mean Reestimation	89.9
VTS–Noise Mean Reestimation	90.6
Proposed–No Noise Mean Reestimation	92.8
Proposed–Noise Mean Reestimation	93.1
Matched	93.9

Table 2: Recognition results (in %) of the Hands-Free (HF) data using different compensation techniques.

6. CONCLUSION

In this paper we first point out that, under a Gaussian mixture assumption an MMSE based compensation algorithm in the log-spectral domain is completely defined by three statistics for each Gaussian component in the Gaussian mixture model. To this end, we propose two new approximations that are well tailored to the mismatch function of [1], and derive the required statistics. We then show how these theoretical results are implemented in a practical compensation algorithm. Results on a digit data base recorded in the car reveal that the proposed method performs significantly better than the baseline, and also the very popular VTS approximation.

7. ACKNOWLEDGEMENT

The author would like to thank Bell Laboratories for availing the corpus, and Dr. Olivier Siohan for his help in obtaining the corpus and earlier collaboration on noise compensation algorithms.

8. REFERENCES

[1] A. Acero, Acoustical and Environmental Robustness in Automatic Speech Recogniiton, Kluwer Academic Publishers, 1993.

[2] M. Afify, O. Siohan, and C.H. Lee, "Upper and Lower Bounds on the Mean of Noisy Speech: Application to Minimax Classification," IEEE Trans. on Speech and Audio Processing, vol. 10, No. 2, February 2002.

[3] P.J. Moreno, B. Raj, and R.M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in Proc. ICASSP, Atlanta, GA, May 1996, pp.733-736.

[4] B. Raj, E. Gouvea, P.J. Moreno, and R.M. Stern, "Cepstral compensation by polynomial approximation for environment-independent speech recognition," in Proc. ICSLP, Philadelphia, PA, Oct 1996.

[5] M.J.F. Gales and S.J. Young, "Cepstral parameter compensation for HMM recognition in noise," Speech Communication, 12:231–239, 1993.