Effect of Noise-in-Speech on MFCC Parameters

LAXMI NARAYANA M
TCS Innovation Lab - Mumbai
Tata Consultancy Services
Yantra Park, Thane (West), Maharastra
INDIA
m.laxminarayana@gmail.com

SUNIL KUMAR KOPPARAPU
TCS Innovation Lab - Mumbai
Tata Consultancy Services
Yantra Park, Thane (West), Maharastra
INDIA
sunilkumar.kopparapu@tcs.com

Abstract: Manifestation of noise-in-speech is inherent unless a conscious effort is made to minimize the disturbances in the surroundings while recording speech. The performance of a speech recognition system often degrades in the presence of noise. In this paper, we study the effect of noise in the speech signal on the extracted speech features that are used in speech recognition. Mel frequency cepstral coefficients (MFCCs) are the most popularly used speech features in speech and speaker recognition applications. We first show theoretically, how additive Gaussian noise with mean, μ and variance, σ^2 effects the speech parameters (MFCCs). The mean and variance of the error in MFCC due to noise-in-speech is related to the mean and variance of the noise added. We experimentally verify that additive Gaussian noise-in-speech results in an error in MFCC parameter estimation which is also Gaussian.

Key–Words: Noisy Speech, MFCC, Error Analysis, Gaussian Noise

1 Introduction

To a large extent, the success of speech recognition and speaker recognition systems relies on their ability to perform robust recognition in noisy environmental conditions. Achieving robustness of speaker recognition and speech recognition systems in (a) the mismatched train-test conditions and (b) noisy environments still remains an unresolved issue today. Although speaker identification in noise-free and matched train-test conditions is almost 100% accurate, the error rate increases drastically when any of these conditions are not met [14]. Performance of speaker identification systems can be improved by either fine-tuning the feature extraction module or the classifier module.

Ming et. al. [7] investigate the problem of speaker identification and verification in noisy conditions. They assume that speech signals are corrupted by environmental noise, but the exact knowledge about the characteristics of noise is assumed to be unknown. They also establish that one single set of features is not optimal across various environmental conditions.

Most often, cepstral features are the speech features of choice for many speaker and speech recognition systems. For example, the Mel-frequency cepstral coefficient (MFCC) representation of speech is probably the oldest [2] most commonly used representation in speaker recognition and and speech recognition [13, 12, 4, 15, 8, 6, 3, 10, 16].

In general, cepstral features are more compact, discriminable, and most importantly, nearly decorrelated such that they allow the diagonal covariance to be used by the hidden Markov models (HMMs) effectively. Therefore, they can usually provide higher baseline performance over filter bank features [5]. The Mel cepstrum has proven to be one of the most successful feature representations in speech related recognition tasks [11] primarily because they are modeled on the understanding of perception of speech by human ear.

In this paper we study the effect of noise-in-speech in the parametric space, more specifically, on the MFCC parameters. We assume the noise to be additive Gaussian with certain mean and variance. The noise-in-speech signal introduces *noise* in the feature extraction process. We derive an expression to show how the noise-in-speech effects the extracted MFCC parameters theoretically and show that the distribution of error in the extracted parameters is related to the additive noise-in-speech statistics through experiments. The rest of the paper is organized as follows. Section 2 derives the error introduced in the computation of MFCC values in the presence of noise in the speech ignal. The derivation uses the MFCC parametric equations operated on clean and noisy speech signals. Section 3 gives the details of the experiments conducted to find the distribution of the error in the MFCC parameters in relation to the error in the speech signal. We conclude in Section 4.

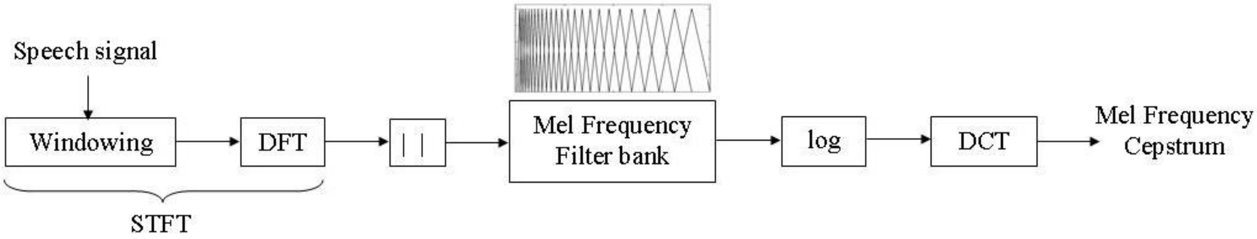

Figure 1: Computation of Mel Frequency Cepstral Coefficients

2 Effect of Noise on the MFCC parameters

The outline of the computation of Mel frequency cepstral coefficients (speech parameters) is shown in Figure 1. In general, the Mel Frequency Cepstral Coefficients (MFCCs) are computed as follows. Let $x[n]$ be a discrete speech signal which is divided into P frames each of length N samples with an overlap of $N/2$ samples such that

$$\{\vec{x}_1[n], \vec{x}_2[n] \cdots \vec{x}_p[n] \cdots \vec{x}_P[n]\}$$

where $\vec{x}_p[n]$ denotes the p^{th} frame of the speech signal $x[n]$ and is

$$\vec{x}_p[n] = \left\{ x\left[p\left(\frac{N}{2} - 1\right) + i \right] \right\}_{i=0}^{N-1} \tag{1}$$

Now the speech signal $x[n]$ can be represented in matrix notation as

$$\hat{X} \overset{def}{=} [\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_p, \cdots, \vec{x}_P]$$

where

$$\vec{x}_p = \begin{bmatrix} x\left[(p-1)\frac{N}{2}\right] \\ x\left[(p-1)\frac{N}{2} + 1\right] \\ \vdots \\ x\left[(p-1)\frac{N}{2} + N - 1\right] \end{bmatrix}$$

Note that the size of the matrix $\hat{X}$ is $N \times P$.

2.1 Windowing and DFT

In speech signal processing, in order to compute the MFCCs of the p^{th} frame, $\vec{x}_p$ is multiplied with a hamming window

$$w[n] = 0.54 - 0.46 \cos\left(\frac{n\pi}{N}\right)$$

followed by the discrete Fourier transform (DFT) as shown in (2).

$$X_p(k) = \sum_{n=0}^{N-1} x_p[n] w[n] \exp^{-j\frac{2\pi kn}{N}} \tag{2}$$

for $k = 0, 1, \cdots, N-1$. If f_s is the sampling rate of the speech signal $x[n]$ then k corresponds to the frequency $f(k) = kf_s/N$. Let $\vec{X}_p = [X_p(0), X_p(1), \cdots, X_p(N-1)]^T$ represent the DFT of $\vec{x}_p$, then, $X = [\vec{X}_1, \vec{X}_2, \cdots \vec{X}_p, \cdots, \vec{X}_P]$ represents the DFT of the windowed p^{th} frame of the speech signal $x[n]$. Note that the size of X is $N \times P$ and is known as STFT (short time Fourier transform) matrix.

2.2 Mel Frequency Filter Bank

The modulus of Fourier transform is extracted and the magnitude spectrum is obtained as $|X|$ which is a matrix of size $N \times P$. The magnitude spectrum is warped according to the Mel scale in order to adapt the frequency resolution to the properties of the human ear [9]. It should be noted that the relation between the Mel frequency and the linear frequency is given by $m_f = 2595 log(1 + f/700)$ [11] where m_f is the Mel frequency and f is the linear frequency. The inverse relationship between f and m_f is given by $f = 700(\exp^{m_f/2595} - 1)$. Then the magnitude spectrum $|X|$ is segmented into a number of critical bands by means of a Mel filter bank which typically consists of a series of, say, F overlapping triangular filters defined by their center frequencies $f_c(m)$. The parameters that define a Mel filter bank are (a) number of Mel filters, (b) minimum frequency and (c) maximum frequency. For speech, in general, it is suggested in [1] that the minimum frequency be greater than 100 Hz. Furthermore, by setting the minimum frequency above 50/60Hz, we get rid of the hum resulting from the AC power, if present. The open source speech recognition engine [1] also suggests that the maximum frequency be less than the Nyquist frequency. Furthermore, there is not much information above 6800 Hz. The Mel filter bank, $M(m,k)$ [17] is given by $M(m,k)$

$$= \begin{cases} 0 & \text{for } f_k < f_c(m-1) \\ \frac{f_k - f_c(m-1)}{f_c(m) - f_c(m) - f_c(m-1)} & \text{for } f_c(m-1) \leq f(k) < f_c(m) \\ \frac{f_k - f_c(m+1)}{f_c(m) - f_c(m) - f_c(m+1)} & \text{for } f_c(m) \leq f(k) < f_c(m+1) \\ 0 & \text{for } f_k \geq f_c(m+1) \end{cases}$$

The Mel filter bank $M(m,k)$ is an $F \times N$ matrix.

2.3 Log Mel Spectrum

The logarithm of the filter bank outputs (Mel spectrum) is given in (3).

$$L(m,p) = ln \left\{ \sum_{k=0}^{N-1} M(m,k)|X(k,p)| \right\} \quad (3)$$

where $m = 1, 2, \cdots, F$ and $p = 1, 2, \cdots, P$. The filter bank output, which is the product of the Mel filter bank, M and the magnitude spectrum, $|X|$ is a $F \times P$ matrix.

2.4 Mel Frequency Cepstrum

A discrete cosine transform of $L(m,p)$ results in the MFCC vector.

$$D(r,p) = \sum_{m=1}^{F} L(m,p) \cos \left\{ \frac{r(2m-1)\pi}{2F} \right\} \quad (4)$$

where $r = 1, 2, \cdots, F$ and $D(r,p)$ is the r^{th} MFCC of the p^{th} frame. The MFCC of all the P frames of the speech signal are obtained as a matrix Φ

$$\Phi\{\hat{X}\} = D = [\Phi_1, \Phi_2, \cdots, \Phi_p, \cdots \Phi_P] \quad (5)$$

Note that the p^{th} column of the matrix Φ represents the MFCC of the speech signal, $x[n]$, corresponding to the p^{th} frame.

2.5 Effect of Noise-in-Speech on MFCC parameters

We now derive how the noise in a speech signal effects the MFCC parameter computation. Let $\gamma[n]$ denote a noise signal having a Gaussian distribution with mean μ and variance σ^2, namely, $\gamma[n] \sim \mathcal{N}(\mu, \sigma^2)$. Let $\gamma_p[n]$ be the p^{th} frame of $\gamma[n]$; extracted in a fashion similar to (1). Note that $\gamma_p[n]$ is Gaussian which has a distribution as shown in (6).

$$\gamma_p[n] \sim \mathcal{N}(\mu_p, \sigma_p^2) = \frac{1}{\sigma_p \sqrt{2\pi}} \exp^{\left\{ \frac{-(n-\mu_p)^2}{2\sigma_p^2} \right\}} \quad (6)$$

In matrix notation, we write $\hat{\gamma} \overset{def}{=} [\vec{\gamma}_1, \vec{\gamma}_2, \cdots, \vec{\gamma}_p, \cdots, \vec{\gamma}_P]$ where

$$\vec{\gamma}_p = \begin{bmatrix} \gamma\left[(p-1)\frac{N}{2}\right] \\ \gamma\left[(p-1)\frac{N}{2}+1\right] \\ \vdots \\ \gamma\left[(p-1)\frac{N}{2}+N-1\right] \end{bmatrix}$$

is the Gaussian noise added to the p^{th} frame of $x[n]$, namely $x_p[n]$. The size of the matrix $\hat{\gamma}$ is of the same size as that of $\hat{X}$, namely, $N \times P$. For convenience, assume $\mu_p = 0$. Then, we have

$$\gamma_p[n] \sim \mathcal{N}(0, \sigma_p^2) = \frac{1}{\sigma_p \sqrt{2\pi}} \exp^{-n^2/2\sigma_p^2}$$

Let $y[n] = x[n] + \gamma[n]$ represent the noisy speech. Accordingly, let $y_p[n] = x_p[n] + \gamma_p[n]$ be the p^{th} frame of the noisy signal. From (2), we can write

$$\begin{aligned} X_p^n(k) &= \sum_{n=0}^{N-1} y_p[n]w[n] \exp^{-j2\pi kn/N} \\ X_p^n(k) &= X_p(k) + \sum_{n=0}^{N-1} \gamma_p[n]w[n] \exp^{-j2\pi kn/N} \\ X_p^n(k) &= X_p(k) + \zeta_p(k) \qquad (7) \end{aligned}$$

The magnitude spectrum of the noisy speech signal is obtained as $|X^n| = |X + \zeta|$ where $\zeta = [\zeta_1, \zeta_2, \cdots, \zeta_p, \cdots, \zeta_P]$. From (4) and (7), we can write the MFCC parameters of noisy speech, $y[n]$ as

$$\begin{aligned} \Phi^n &= \Phi\{y[n]\} \\ &= \sum_{m=1}^{F} L^n(m,p) cos \left\{ \frac{r(2m-1)\pi}{2F} \right\} \qquad (8) \end{aligned}$$

where

$$L^n(m,p) = ln \left\{ \sum_{k=0}^{N-1} M(m,k)|X(k,p) + \zeta(k,p)| \right\} \quad (9)$$

The error in the MFCC parameters, E, due to additive noise γ in speech can be derived from (3), (4), (8) and (9)

$$\begin{aligned} E &= \Phi^n - \Phi \qquad (10) \\ &= \sum_{m=1}^{F} V ln \left\{ \frac{\sum_{k=0}^{N-1} M(m,k)|X(k,p) + \zeta(k,p)|}{\sum_{k=0}^{N-1} M(m,k)|X(k,p)|} \right\} \end{aligned}$$

where

$$V \overset{def}{=} \cos \left\{ \frac{r(2m-1)\pi}{2F} \right\}$$

Figure 2 shows the computation of the error in the MFCC parameters (E). Apply triangular inequality, $|a+b| \leq |a| + |b|$ we get

$$E \leq \sum_{m=1}^{F} V ln \left\{ 1 + \frac{\sum_{k=0}^{N-1} M(m,k)|\zeta(k,p)|}{\sum_{k=0}^{N-1} M(m,k)|X(k,p)|} \right\} \quad (11)$$

The error, E in MFCC due to noise-in-speech is upper bounded by the RHS in (11)[1].

[1] We are in the process of showing that the distribution of E is Gaussian with μ_E and σ_E^2 which is related to the μ and σ^2 of the noise

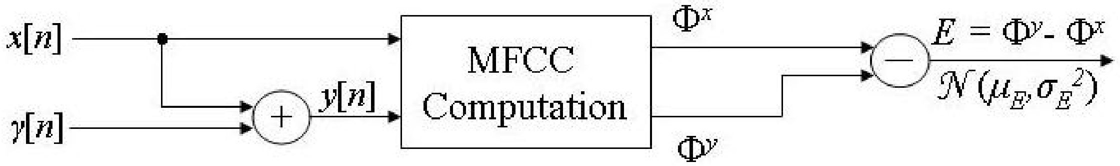

Figure 2: Calculation of the error signal

3 Experimental Results

In all our experiments we consider speech signal sampled at 8 kHz and represented by 16 bits. The speech signal is divided into frames of duration 20 ms (or $N = 160$ samples) and 10 ms overlap (or $N/2 = 80$ samples). MFCC features are computed for each speech frame. Initially the speech frame is pre processed by passing it through a hamming window and the magnitude spectrum of the Fourier transform is warped according to the Mel scale. The Mel filter bank has 31 bands spread from 300 Hz (minimum frequency) to a maximum frequency of 3500 Hz.

Initially the MFCC parameters (Φ) are computed for the clean speech $x[n]$ as described in Section 2 (see (4) of Figure 1). Gaussian noise with varying mean (μ) and variance (σ^2) are generated $\gamma^{(\mu,\sigma^2)}[n]$ and added to the speech signal $x[n]$ to generate noisy speech

$$y^{(\mu,\sigma^2)}[n] = x[n] + \gamma^{(\mu,\sigma^2)}[n]$$

The signal to noise ratio, SNR of the noisy speech $y^{(\mu,\sigma^2)}[n]$ is given by

$$\begin{aligned} SNR &= \frac{P_x}{P_\gamma} = \left(\frac{A_x}{A_\gamma}\right)^2 \\ SNR_{dB} &= 10\log\left(\frac{P_x}{P_\gamma}\right) = 20\log\left(\frac{A_x}{A_\gamma}\right) \end{aligned} \quad (12)$$

where P_x and A_x are the average power and root mean square (RMS) amplitude respectively, of the speech signal, $x[n]$; and P_γ and A_γ are the average power and RMS amplitude respectively, of the Gaussian noise $\gamma^{(\mu,\sigma^2)}[n]$. MFCCs that are extracted from $y^{(\mu,\sigma^2)}[n]$ are denoted by Φ^y. The error in the MFCC parameters due to additive noise is given by $E = \Phi^y - \Phi$. The mean, μ_E and variance, σ_E^2 of the error in MFCC estimation[2], for varying additive Gaussian noise, namely $\gamma^{(\mu,\sigma^2)}[n]$ and the corresponding SNR are tabulated in Table 1.

It can be observed from Table 1 that μ of the additive noise does not effect the estimation of MFCC parameters, this can be observed from Table 1 where μ_E (third column) varies minimally with variation in μ (column 1 in Table 1). Infact even a large variation of $\mu(= 5)$ has minimal influence on μ_E. However, as expected σ_E^2 increases proportionately with an increase in σ^2 and is almost constant when σ^2 is constant (see column 3 in Table 1 for the rows corresponding to constant σ^2).

Table 1: Error in MFCC with varying noise-in-speech

μ	σ^2	μ_E	σ_E^2	SNR	
Trivial	0	0	0	0	–
Constant μ	0	1	0.0545	3.967	64.3
	0	2	0.0619	5.028	57.5
	0	3	0.0668	5.703	53.5
	0	4	0.0686	6.224	50.6
	0	5	0.0718	6.657	48.3
Constant σ^2	1	1	0.0604	3.956	57.5
	2	1	0.0665	3.960	48.4
	3	1	0.0727	3.941	41.4
	4	1	0.0781	3.943	36.1
	5	1	0.0831	3.944	31.9

The relation between the variance of the error in calculating MFCC (σ_E^2) and the variance of the additive Gaussian noise (σ^2) can be determined as a first order polynomial fit, namely,

$$\sigma_E^2 \approx 1.16\sigma^2 + 1.7 \quad (13)$$

4 Conclusion

The performance of a speech recognition system often degrades in noisy conditions. Mel frequency cepstral coefficients are the popularly used speech features in speech and speaker recognition systems. The effect of additive Gaussian noise-in-speech on the extracted MFCC parameters is studied. We determined the upper bound on the error in extracting the MFCC parameters of noisy speech theoretically and have shown experimentally that the variance in the error of extracted MFCC parameters is related to the variance of the additive Gaussian noise. Experimental results show that the mean of the additive Gaussian noise does not have much influence on the parameters of the error. The

[2] We assume that distribution of the error in MFCC is Gaussian

variance of the error varies as a function of the additive Gaussian noise variance. We are in the process of showing theoretically that the distribution of the error in computing MFCC parameters is indeed Gaussian and the variance in the error in MFCC parameters is related to the variance of the additive noise.

References:

[1] CMU. http:// cmusphinx.sourceforge.net/ sphinx4/ javadoc/ edu/ cmu/ sphinx/ frontend/ frequencywarp/ melfrequencyfilterbank.html.

[2] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Processing*, 28, no. 4:357–366, 1980.

[3] G. Friedland and O. Vinyals. Live speaker identification in conversations. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 1017–1018, New York, NY, USA, 2008. ACM.

[4] M. R. Hasan, M. Jamil, M. G. Rabbani, and M. S. Rahman. Speaker identification using mel frequency cepstral coefficients. *3rd International Conference on Electrical & Computer Engineering ICECE 2004*, 28-30 December 2004, Dhaka, Bangladesh.

[5] Z. Jun, S. Kwong, W. Gang, and Q. Hong. Using mel-frequency cepstral coefficients in missing data technique. *EURASIP Journal on Applied Signal Processing*, 2004, no. 3:340–346, 2004.

[6] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Mobisys '09: Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 165–178, New York, NY, USA, 2009. ACM.

[7] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds. Robust speaker recognition in noisy conditions. *IEEE Trans. Audio, Speech and Language Processing*, 15:1711–1723, 2007.

[8] H. Moeinzadeh, M.-M. Mohammadi, A. Akbari, and B. Nasersharif. Robust speech recognition using evolutionary class-dependent LDA. In *GECCO '09: Proceedings of the 11th annual conference companion on Genetic and evolutionary computation conference*, pages 2109–2114, New York, NY, USA, 2009. ACM.

[9] S. Molau, M. Pitz, R. S. Uter, and H. Ney. Computing mel-frequency cepstral coefficients on the power spectrum. *Proc. Int. Conf. on Acoustic, Speech and Signal Processing*, pages 73 – 76, 2001.

[10] N. S. Nehe and R. S. Holambe. New robust subband cepstral feature for isolated world recognition. In *ICAC3 '09: Proceedings of the International Conference on Advances in Computing, Communication and Control*, pages 326–330, New York, NY, USA, 2009. ACM.

[11] T. F. Quatieri. Discrete-time speech signal processing: Principles and practice. *Pearson Education*, II:686, 713, 1989.

[12] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17, No. 1-2:91–108, 1995.

[13] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3, No. 1, January 1995.

[14] R.Padmanabhan and H. A. Murthy. Feature evaluation for speaker identification in radio communications channel. *Proceedings of The Fifteenth National Conference on Communications, NCC 2009*, pages 307–310, January 16 - 18, 2009, IIT Guwahati, India.

[15] H. Seddik, A. Rahmouni, and M. Sayadi. Text independent speaker recognition using the mel frequency cepstral coefficients and a neural network classifier. *First International Symposium on Control, Communications and Signal Processing*, pages 631–634, 2004.

[16] J. Shen, J. Shepherd, B. Cui, and K.-L. Tan. A novel framework for efficient automated singer identification in large music databases. *ACM Trans. Inf. Syst.*, 27(3):1–31, 2009.

[17] S. Sigurdsson, K. B. Petersen, and T. L. Schiler. Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. *Conference Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR)*, Vicoria, Canada, 2006.