The Effect of Additive Noise on Speech Amplitude Spectra: A Quantitative Analysis

Qifeng Zhu, *Member, IEEE*, and Abeer Alwan, *Senior Member, IEEE*

***Abstract*—This letter analyzes the effect of additive noise on speech amplitude spectra, and introduces a method to estimate speech spectra from noisy observations. Estimated spectra are used to compute the Mel-Frequency Cepstral Coefficients as a recognition front-end. Compared to linear spectral subtraction, this technique improves the performance of digit recognition in noise.**

***Index Terms*—Robust speech recognition, spectral subtraction.**

I. Introduction

NOISE ROBUSTNESS is an important challenge for automatic speech recognition (ASR). It is typically handled by the acoustic model [often a hidden Markov model (HMM)], and/or at the front-end (feature extraction).

When the noise is additive and stationary, and if one can estimate the average noise spectrum, a widely used technique for noise removal is linear spectral subtraction (SS) [1], [2]. SS attempts to remove noise effects by subtracting the average magnitude spectrum of the noise $|\overline{N}(e^{j\omega})|$ from that of the observed signal $|X(e^{j\omega})|$:

$$\left|\tilde{S}(e^{j\omega})\right| = |X(e^{j\omega})| - \left|\overline{N}(e^{j\omega})\right| \tag{1}$$

where $|\tilde{S}(e^{j\omega})|$ is the estimate of the speech spectrum. Clearly, this is not valid when the signal and noise are not in phase [3].

Nonlinear spectral subtraction (NSS) [4] is an alternative method which can improve ASR noise robustness by using a subtraction factor that is a function of SNR:

$$\left|\tilde{S}(e^{j\omega})\right| = |X(e^{j\omega})| - \frac{\left|\overline{N}(e^{j\omega})\right|}{1+\gamma\cdot \mathrm{SNR}(e^{j\omega})} \tag{2}$$

where γ is a tunable parameter.

This letter studies the effect of additive noise on speech amplitude spectra in a more quantitative way. A speech spectral recovery method is proposed based on the analysis.

II. Additive Noise Model

Consider a speech signal $s(k)$ which is affected by additive noise $n(k)$. The observed signal in the frequency domain can be expressed as $X(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega})$.

Manuscript received March 28, 2002; revised April 4, 2002. This work was supported in part by the National Science Foundation, STM, Broadcom, HRL, and Mindspeed together with the state of California through the University of California MICRO program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yair Shoham.

The authors are with The Henry Samueli School of Engineering and Applied Science, University of California, Los Angeles, CA 90095-1591 USA (e-mail: qifeng@icsl.ucla.edu; alwan@icsl.ucla.edu).

Digital Object Identifier 10.1109/LSP.2002.801722.

Because magnitude spectra are often used in computing ASR features such as the Mel-Frequency Cepstral Coefficients (MFCCs), it is important to study how magnitude spectra are affected by noise.

Consider a frequency component ω_0. The speech spectrum at that frequency can be regarded as an unknown signal, and the noise as a random variable. The complex spectra of the noise and speech at frequency ω_0 can be written as

$$N(e^{j\omega_0}) = be^{j\theta_2}$$
$$S(e^{j\omega_0}) = ae^{j\theta_1}.$$

Then

$$|X(e^{j\omega_0})| = p = |ae^{j\theta_1} + be^{j\theta_2}|.$$

For the noise spectrum, both b and θ_2 are random variables; p is the amplitude of the observed spectrum $X(e^{j\omega})$ at ω_0, which is also a random variable. Depending on the relationship between θ_1 and θ_2, p may be larger or smaller than the signal amplitude a. The expectation of p can be computed with respect to b and θ_2.

Assume that θ_2 is uniformly distributed between 0–2π. Hence the expectation of p is

$$\begin{aligned} E\{p\} &= \frac{1}{2\pi}\int_b f(b)\cdot\int_0^{2\pi} |ae^{j\theta_1} + be^{j\theta_2}|\, d\theta_2\, db \\ &= \frac{1}{2\pi}\int_b f(b)\cdot\int_0^{2\pi} \sqrt{a^2+b^2+2ab\cos(\theta)}\, d\theta\, db \end{aligned} \tag{3}$$

where $f(b)$ is the probability distribution function of b, which is typically Laplacian, and $\theta = \theta_2 - \theta_1$ (the relative phase between the signal and noise spectrum at ω_0).

We use the approximation $E(F(b)) \approx F(E(b)) = F(\overline{b})$, where $F(b)$ is the integral with respect to θ. The accuracy of this approximation depends on the distribution of b and the values of $Q(b)$ around b. The smaller the variance of b, the more accurate is this approximation. Thus

$$\begin{aligned} E\{p\} &\approx \frac{1}{2\pi}\int_0^{2\pi} \sqrt{a^2+\overline{b}^2+2a\overline{b}\cos(\theta)}\, d\theta \\ &= \frac{\overline{b}}{2\pi}\int_0^{2\pi} \sqrt{1+r^2+2r\cos(\theta)}\, d\theta \\ &= \overline{b}\cdot Q(r) \end{aligned} \tag{4}$$

where

$$r = a/\overline{b}$$
$$Q(r) = \frac{1}{2\pi}\int_0^{2\pi} \sqrt{1+r^2+2r\cos(\theta)}\, d\theta.$$

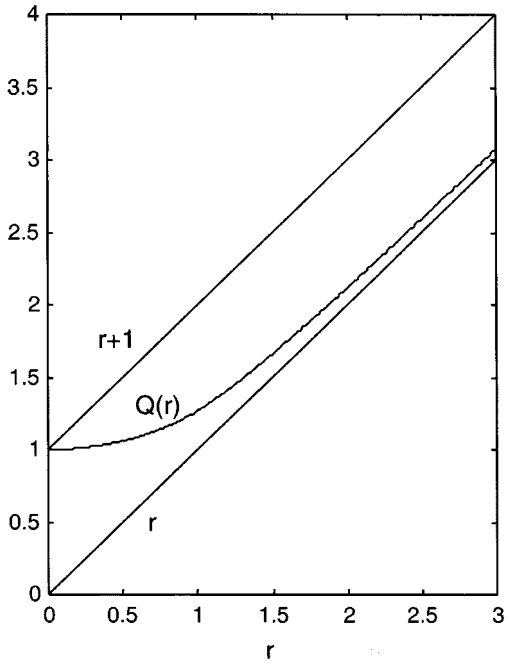

Fig. 1. Numerical solution of $Q(r)$ in the range of r from 0 to 3.

There is no closed-form solution for $Q(r)$. Numerical solutions, however, can be computed. Fig. 1 shows $Q(r)$ computed numerically as a function of r. $Q(r)$ lies between $r+1$ and r; if $Q(r)=r+1$, then $E\{p\}=a+\overline{b}$, which is the relation used in SS [in (1)]. However, when r is large (i.e., signal amplitude is much higher than the noise), $Q(r)$ is nearly equal to r (hence $E\{p\}=a$), or the signal amplitude is not affected much by noise. When r is small, on the other hand, $Q(r)$ is close to 1, which means that the observed amplitude is nearly the same as the average noise amplitude.

Note that this analysis supports the intuition behind NSS in that the estimated signal amplitude is close to the observed amplitude for high SNRs (minimal subtraction is needed), while NSS is close to linear spectral subtraction when the SNR is low. This letter provides a quantitative framework of the problem.

III. Algorithm and Implementation of Spectral Recovery

If we know $E\{p\}$ and $\overline{b}$, we can recover the signal amplitude a using (4). One possibility is to approximate $Q(r)$ by a third-order polynomial $T(r)$ using mean square error fitting: $Q(r)\approx T(r)=0.9820-0.0075r+0.3761r^2-0.0461r^3$ for $0<r<2$. For large r ($r>2$), $Q(r)\approx r$. Often we do not know $E\{p\}$, so we have to replace the expectation with the instant observation p. The signal amplitude a can be recovered using the following equation:

$$\begin{aligned} T(r)\approx Q(r)=E\{p\}/\overline{b}\approx p/\overline{b} \\ \Rightarrow r=T^{-1}\left(p/\overline{b}\right) \\ a=\overline{b}r. \end{aligned} \tag{5}$$

This equation has three roots, but only one is in the range of interest $(0<r<2)$.

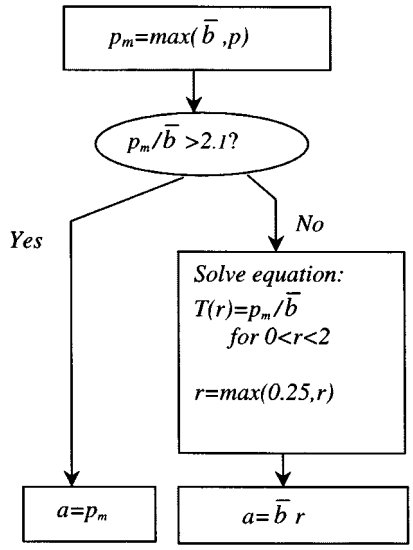

Fig. 2. Flow diagram of spectral recovery.

For each frequency component, with the observed amplitude p and the average noise amplitude $\overline{b}$, the recovered speech spectral amplitude a is obtained as shown in Fig. 2. Equation solving can be realized with a computationally efficient lookup table technique. Note that using p instead of $E\{p\}$ can lead to a biased estimate of a through the nonlinear equation. This bias depends on the distribution of p and $Q(r)$.

The maximum operation in the first step is to prevent the observed amplitude p from being smaller than the average noise amplitude $\overline{b}$; r is close to $Q(r)$ when $Q(r)>2.1$, e.g., $Q(2)=2.1$. A noise floor for r is selected as 0.25 because, when the noise is much higher than the observed signal ($r<0.25$), the speech estimate is not precise.

IV. Experiments and Results

The goal of the ASR experiments is to verify the accuracy of the analysis in controlled conditions. The experiments utilized a HMM-based system (HTK2.2) for an isolated digit recognition task using the TI46 database. For each digit, one HMM with four states and two mixtures is trained from 160 utterances spoken by 16 talkers (eight males and eight females). Training includes two steps of Viterbi and forward/backward training with four iterations each. A Viterbi algorithm is used for recognition using 480 different utterances. Training is done with clean signals and recognition with noisy signals at different SNRs.

Silence is removed from every digit file and computer-generated speech-shaped noise is applied. Preemphasis with a first-order finite-impulse response (FIR) filter (coefficient of -0.95) is used. MFCCs are computed with a frame length of 25 ms and frame step size of 10 ms. The average noise spectrum after preemphasis is precomputed and used for spectral recovery.

For comparison, linear spectral subtraction is implemented in the same way as in [2]: $|\hat{S}(e^{j\omega})|=\max(|\tilde{S}(e^{j\omega})|,\beta|\overline{N}(e^{j\omega})|)$, where β is a factor for the noise floor, and $|\hat{S}(e^{j\omega})|$ is the estimated speech spectrum. We use $\beta=0.25$, which is the optimal number for the SS method, and is the same as the noise floor

TABLE I
DIGIT RECOGNITION ACCURACY (IN PERCENT) USING MFCCS, MFCCS WITH SPECTRAL RECOVERY, MFCCS WITH SS, AND MFCCS WITH NSS

SNR	0 dB	3 dB	5 dB	10 dB	20 dB
MFCC	25.00	40.21	52.29	85.83	98.96
MFCC + Eq. 5	38.12	54.17	66.88	89.17	98.12
MFCC + SS	30.62	48.75	65.42	86.25	97.71
MFCC + NSS	36.25	55.62	68.75	90.83	98.33

parameter in our spectral recovery technique. Thus, for both implementations the estimated speech spectrum is not lower than 0.25 times the average noise spectrum.

Table I shows recognition results. The first row shows results with MFCCs only. The second row shows the results of spectral recovery. The third row shows results with SS, and the last row shows results with NSS for $\gamma = 0.8$ (tuned to maximize recognition accuracy). Clearly, spectral recovery improves ASR results over SS without parameter tuning. Note that NSS, which uses the same qualitative idea and with parameter tuning, could lead to results similar to spectral recovery.

In the experiments, $\overline{b}$ is assumed to be known. In practice, however, $\overline{b}$ could be estimated with a speech/nonspeech detection technique. When $p \gg \overline{b}$, the estimate of a is not sensitive to that of $\overline{b}$, because the estimate of a will be close to p. But when p is close to $\overline{b}$, the estimate of a becomes quite sensitive to $\overline{b}$. This suggests that when the signal is buried by noise, it is difficult to recover the signal amplitude using $\overline{b}$ and p.

V. CONCLUSION AND DISCUSSION

In this letter, a quantitative analysis of how the speech amplitude spectrum is affected by noise is introduced. It supports the qualitative idea in NSS to apply a minimum subtraction factor at high SNRs and subtract more noise at low SNRs. A method to recover the speech amplitude from the observed amplitude spectrum given the average noise amplitude spectrum is derived, with no parameter tuning. Compared with linear spectral subtraction, this method improves the performance of speech recognition systems in noise.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 126–137, Mar. 1999.

[3] I. Y. Soon, S. Koh, and C. Yeo, "Selective magnitude subtraction for speech enhancement," in *Proc. Fourth Int. Conf. High Performance Computing Asia-Pacific Region*, vol. 2, 2000, pp. 692–695.

[4] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (nss), hidden Markov models and the projection or robust speech recognition in cars," *Speech Commun.*, vol. 11, no. 2–3, pp. 215–228, 1992.