SIGNAL PROCESSING METHODS FOR REMOVING THE EFFECTS OF WHOLE-BODY VIBRATION UPON SPEECH

Rachel M. Bittner

New York University
Music and Audio Research Laboratory
35 W. 4th St, New York, NY 10003, USA
rmb456@nyu.edu

Durand R. Begault

NASA Ames Research Center
Human Systems Integration Division
Moffett Field, CA, 94035, USA
durand.r.begault@nasa.gov

ABSTRACT

Humans may be exposed to whole-body vibration in environments where clear speech communications are crucial, particularly during the launch phases of space flight and in high-performance aircraft. Prior research has shown that high levels of vibration cause a decrease in speech intelligibility. However, the effects of whole-body vibration upon speech are not well understood, and no attempt has been made to restore speech distorted by whole-body vibration. In this paper, a model for speech under whole-body vibration is proposed and a method to remove its effect is described. The method described reduces the perceptual effects of vibration, yields higher ASR accuracy scores, and may significantly improve intelligibility. Possible applications include incorporation within communication systems to improve radio-communication systems in environments such a spaceflight, aviation, or off-road vehicle operations.

Index Terms— Whole-Body Vibration, Speech Intelligibility

1. INTRODUCTION

Speech production is inhibited when humans are exposed to whole-body vibration between 2 and 20 Hz [1]. Examples of environments where humans are exposed to these vibration levels include spacecraft, high-performance aircraft, military land vehicles, and heavy machinery such as tractors. In these contexts clear speech communications are crucial; in particular, speech intelligibility for radio communications between crew and ground control is of concern during launch phases of space flight because other means of communication such as operation of manual controls are extremely difficult if not impossible. NASA standards require speech intelligibility levels to be equivalent to a 90% word identification [2], but prior research has shown that speech under whole-body vibration is at least 9% less intelligible than speech in non-vibrated conditions [3]. Even in situations where intelligibility remains high, "distortions of the speech signal will increase listening effort and fatigue, and reduce speech quality to the point where communication becomes difficult and annoying" [4].

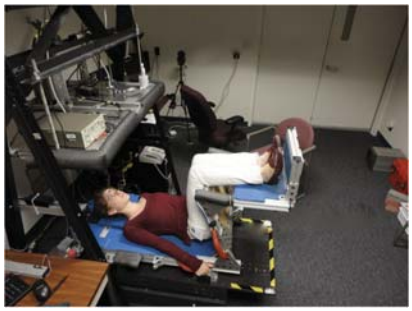

Fig. 1. Speaker positioned in the semi-supine position on an experimental vibration platform.

NASA has addressed the need for developing analytic models for human vibration response in order to predict the effects on manual performance and speech production [4]. Previous studies [5, 6, 7, 8, 9] have examined the physical effects of whole-body vibration on mechanisms of the vocal production system and examined the distortion of the speech signal. These studies found that vibration between 2 and 20 Hz causes disruptions in airflow which in turn cause frequency and amplitude modulations in the resulting speech. However, no model for speech under whole-body vibration has been proposed, and no effort has been made to address this reduction in intelligibility. It is generally difficult or impossible to remove the vibration itself, warranting methods to improve speech intelligibility that do not involve changing the vibration environment. There has been no previous research on removing the vibration effect from the speech signal directly.

Whole-body vibration is defined by Griffin [1] as occurring "when the body is supported on a surface which is vibrating", such as when sitting on a vibrating seat, standing on a vibrating floor, or lying on a vibrating bed. The studies addressed in this paper consider speakers positioned in the face-up recumbent (semi-supine) position affected by sinusoidal

vibration in the body's x-axis (back to chest) as shown in Figure 1. We examine sinusoidal vibrations with constant frequency because they estimate the vibration present in spaceflight environments. Sinusoidal vibration levels will be characterized in this paper by frequency (Hz) and 0-peak acceleration amplitude (measured in units of earth's gravity g). We focus on the communication channel between a speaking crew member (exposed to vibration) and a listener in a ground control scenario (not exposed to vibration). Unlike more common noise reduction problems, the goal is not to remove background artifacts, but to remove distortion from the source itself. This paper proposes a model for vibrated speech and presents a method to remove or reduce the effects of vibration on the speech signal to improve speech quality and intelligibility.

2. A MODEL FOR VIBRATED SPEECH

A study similar in setup to [9] was conducted in the Human Vibration Laboratory at NASA Ames Research Center. Speech samples of sustained phonemes and sentences were gathered from 6 speakers at 4 vibration conditions[1]. The model proposed here is motivated by analysis of this data and results from similar studies such as [5, 6, 7, 9].

The primary observed characteristics of the data was that the fundamental frequency, energy, and formant frequencies of vibrated speech oscillate as a function of the vibration acceleration. An example of these characteristics can be seen in Figure 2. Our model for vibrated speech is based upon the source-system model for speech production, where each short time frame $s_{\hat{n}}[n]$ of a speech signal $s[n]$ is modeled as the output of a filter with coefficients $\boldsymbol{\alpha}_{\hat{n}} = \{\alpha_{\hat{n}}(k)\}_{k=1}^{p}$ and source "excitation" $e_{\hat{n}}[n]$. Pitch and energy oscillations are modeled as modulations of the excitation $e[n]$, and formant frequency oscillations are modeled as modulations of the filter coefficients $\boldsymbol{\alpha}$. Under the assumption that the vibration is sinusoidal with known constant frequency f_{v}, we assume the source excitation is amplitude modulated by the function

$$M_a(t) = A\sin(2\pi f_{\text{v}}(t+k)) + B \tag{1}$$

and frequency modulated by

$$M_f(t) = t - \frac{D}{2\pi f_{\text{v}}}\cos(2\pi f_{\text{v}}(t+h)) \tag{2}$$

such that $0 < A < B$, $0 < D \leq 1$, and $h, k \in \left[-\frac{1}{2f_{\text{v}}}, \frac{1}{2f_{\text{v}}}\right]$.

The models for $M_a(t)$ and $M_f(t)$ are based on the premise that the airflow quantity passing through the vocal tract during whole-body vibration is proportional to the acceleration acting on the body. Oscillatory quantities of airflow passing through the vocal tract cause an effect (similar to musical vibrato) in which energy and frequency of the voice

[1] 8 Hz, 0.5g; 12 Hz 0.5g; 12 Hz, 0.7g; and 16 Hz 0.5g

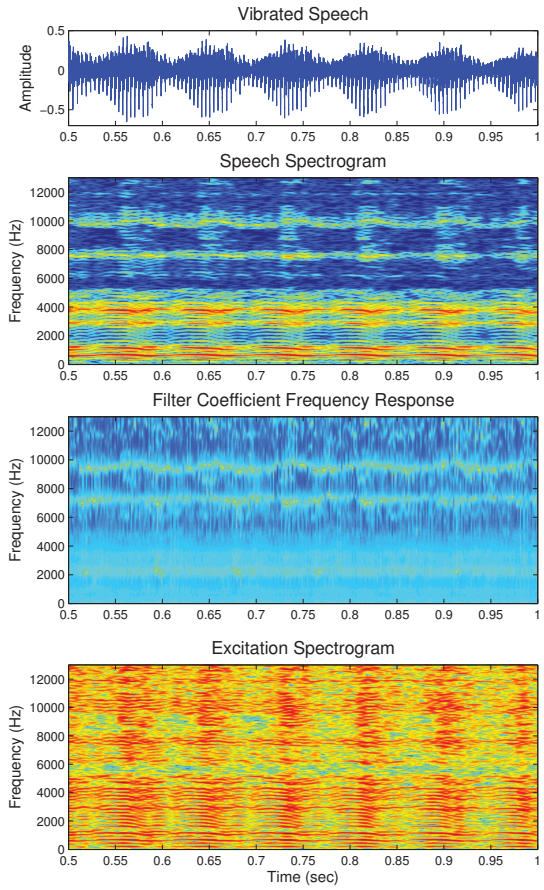

Fig. 2. Effect of vibration on a sustained vowel [o] at 12 Hz vibration. The waveform and spectrogram of the vibrated vowel (top, upper middle) are shown, along with the frequency response of the source-system filter over time (lower middle) and the spectrogram of the excitation $e[n]$ (bottom). Sustained vibrated phonemes sound somewhere between a very wide musical vibrato and a bleating goat.

oscillates along with the airflow. Note that this formulation is different from traditional AM/FM modulation in the sense that the roles of the carrier and modulator are reversed.

The analog source excitation $\xi(t)$ is decomposed as a sum of sinusoids as in [10], such that

$$\xi(t) = \sum_{i=1}^{L} a_i \sin(2\pi f_i t + \phi_i) \tag{3}$$

where the parameters a_i, f_i, ϕ_i are respectively the amplitude, frequency and phase of sinusoid i, and L is the number of sinusoids in the decomposition. The resulting vibrated excitation $\tilde{e}[n]$ is given by

$$\tilde{e}[n] = M_a(t_n) \cdot E\left(M_f(t_n)\right) \tag{4}$$

where t_n is a sequence of time samples. Finally, the model

for vibrated speech can be written as

$$\tilde{s}_{\hat{n}}[n] = \sum_{k=1}^{p} \tilde{\boldsymbol{\alpha}}_{\hat{n}}(k)\tilde{s}_{\hat{n}}[n-k] + \tilde{e}_{\hat{n}}[n] \tag{5}$$

where $\tilde{\boldsymbol{\alpha}}_{\hat{n}}$ is a vector of modulated filter coefficients.

2.1. Parameter Estimation

Given an observed vibrated excitation, the parameters for the model can be estimated. The amplitude modulation parameters are chosen by fitting the observed data to the vibrated excitation model in Equation (4). However, the frequency-modulated source excitation $\xi(M_f(t_n))$ is complex. $\xi(M_f(t_n))$ is simplified for this stage of parameter estimation to be a random process $w[n]$, such that for each n the expected value $E\left(w[n]\right) = 0$ and $E\left(w[n]^2\right) = 1$. Then the vibrated excitation can be approximated by

$$\tilde{e}[n] \approx (A\sin(2\pi f_{\text{v}}(t_n + k)) + B) \cdot w[n] \tag{6}$$

Let $y[n]$ be an observed vibrated excitation. The parameters A, B, and k are chosen to minimize the expected value of the sum of the squared distances between $|y[n]|$ and $|\tilde{e}[n]|$. If the sum is taken over $N = j\frac{f_s}{f_{\text{v}}}$ points where $j \in \mathbb{N}$ and f_s is the sampling frequency, the parameters can be solved for analytically. Up to a multiplicative constant, the optimal parameters are

$$A = \frac{2}{N}\sqrt{\left(\sum_{n=1}^{N} |y[n]|\sin(2\pi f_{\text{v}} t_n)\right)^2 + \left(\sum_{n=1}^{N} |y[n]|\cos(2\pi f_{\text{v}} t_n)\right)^2}$$

$$B = \frac{1}{N}\sum_{n=1}^{N} |y[n]|$$

$$k = \frac{1}{2\pi f_{\text{v}}}\arctan\left(\frac{\sum_{n=1}^{N} |y[n]|\cos(2\pi f_{\text{v}} t_n)}{\sum_{n=1}^{N} |y[n]|\sin(2\pi f_{\text{v}} t_n)}\right)$$

The frequency modulation parameters are chosen by fitting the amplitude de-modulated signal to the frequency modulated excitation model:

$$\xi(M_f(t_n)) = \sum_{i=1}^{L} a_i \sin\left(2\pi f_i M_f(t_n) + \phi_i\right) \tag{7}$$

Instead of fitting the model to the data directly, time/frequency tracks in the STFT (sequences of neighboring local maxima over time) as described in [10] of the data are fit to the instantaneous frequency of a sinusoid in Equation (7). The instantaneous phase of sinusoid j is given by

$$\varphi_j(t_n) = 2\pi f_j t_n - \frac{D f_j}{f_{\text{v}}}\cos(2\pi f_{\text{v}}(t_n + h)) + \phi_j \tag{8}$$

and the corresponding instantaneous frequency is

$$\varphi_j'(t_n) = 2\pi f_j + 2\pi D f_j \sin(2\pi f_{\text{v}}(t_n + h)) \tag{9}$$

For each time/frequency track $\omega_j[n]$, an initial estimate of the parameters D_j, f_j, and h_j are chosen to minimize the sum of the squared errors between $\omega_j[n]$ and the modeled instantaneous frequency $\varphi_j'(t_n)$. If the sum is taken over N points as described in the amplitude modulation case, the parameters can also be solved for analytically, giving the optimal parameters

$$D_j = \frac{\sqrt{\left(\sum_{n=1}^{N} \omega_j[n]\sin(2\pi f_{\text{v}} t_n)\right)^2 + \left(\sum_{n=1}^{N} \omega_j[n]\cos(2\pi f_{\text{v}} t_n)\right)^2}}{\frac{1}{2}\sum_{n=1}^{N} \omega_j[n]}$$

$$f_j = \frac{1}{2\pi N}\sum_{n=1}^{N} \omega_j[n]$$

$$h_j = \frac{1}{2\pi f_{\text{v}}}\arctan\left(\frac{\sum_{n=1}^{N} \omega_i[n]\cos(2\pi f_{\text{v}} t_n)}{\sum_{n=1}^{N} \omega_i[n]\sin(2\pi f_{\text{v}} t_n)}\right)$$

The final parameters are chosen from the estimate j^* with the smallest overall error, such that $D = D_{j^*}$ and $h = h_{j^*}$.

3. REMOVING VIBRATION

The signal is preprocessed by first high-pass filtering with a cutoff at 40 Hz to remove additive mechanical noise from the vibrating platform. Next the speech is filtered using a typical speech pre-emphasis filter, which balances the low and high frequencies in speech for more accurate analysis. Finally, the speech is separated into evenly-spaced 6 ms frames and the frames are grouped by phoneme. The following steps are performed to remove vibration from phonemes sustained for at least one period of vibration: (1) perform frame by frame linear predictive analysis to extract filter coefficients and excitation; (2) estimate amplitude modulation parameters from vibrated excitation and remove amplitude modulation; (3) estimate frequency modulation parameters and remove frequency modulation; (4) compute smoothed filter coefficients; (5) generate recovered speech using smoothed filter coefficients and source excitation. These steps are described in detail below.

For each time frame $\tilde{s}_{\hat{n}}[n]$, the coefficient vector $\tilde{\boldsymbol{\alpha}}$ and the excitation $\tilde{e}_{\hat{n}}[n]$ are computed using linear predictive analysis. The frames $\tilde{e}_{\hat{n}}[n]$ are combined using overlap-add to form the excitation signal $\tilde{e}[n]$.

The amplitude modulation model parameters k, A, and B are computed as described in Section 2.1. Given these parameters, the amplitude modulation can be removed from the vibrated excitation, leaving only the frequency modulated source excitation:

$$\xi(M_f(t_n)) = \frac{\tilde{e}[n]}{A\sin(2\pi f_{\text{v}}(t_n + k)) + B} \tag{10}$$

The frequency modulation model parameters D and h are then computed from $\xi(M_f(t_n))$ as described in Section 2.1. The frequency modulation is removed by frequency modulating $\xi(M_f(t_n))$ by the function $\dfrac{1}{2\pi D\sin\left(2\pi f_{\text{v}}(t_n + h)\right) + 1}$ via resampling the signal in short time intervals.

Given the recovered source excitation $e[n]$ and the vibrated coefficients $\tilde{\boldsymbol{\alpha}}(\hat{n})$ from the original short time frames,

Speaker	Clean	Vibrated	De-vibrated
1	69.3%	51.2%	54.7%
2	66.5%	50.4%	52.6%
3	58.0%	38.1%	40.6%
4	58.7%	46.3%	49.7%
5	50.2%	45.1%	47.4%
6	59.4%	38.4%	47.0%
Mean	60.4%	44.9%	48.7%

Table 1. Vowel classification accuracy for each speaker.

a partially de-vibrated speech signal $\bar{s}[n]$ is generated

$$\bar{s}_{\hat{n}}[n] = \sum_{k=1}^{p} \tilde{\boldsymbol{\alpha}}_{\hat{n}}(k)\bar{s}_{\hat{n}}[n-k] + e_{\hat{n}}[n] \tag{11}$$

The "true" filter coefficients $\boldsymbol{\alpha}_{\hat{n}}$ are estimated by performing another round of linear predictive analysis on $\bar{s}[n]$ for time frames $\hat{m}$ with length equal to the vibration period $\frac{1}{f_v}$. Each resulting coefficient vector $\bar{\boldsymbol{\alpha}}_{\hat{m}}$ gives an estimate of a sequence of the "true" filter coefficients $\boldsymbol{\alpha}_{\hat{n}}$. If $\hat{n}_i = \frac{T_S}{2} + iT_S$, and $\hat{m}_j = \frac{T_L}{2} + jT_L$, where T_S and T_L are the lengths of the short and long time frames respectively, then $\bar{\boldsymbol{\alpha}}_{\hat{m}_j} \approx \boldsymbol{\alpha}_{\hat{n}_i}$ for all i such that $jT_L \leq \hat{n}_i < (j+1)T_L$. The final recovered speech is given by:

$$s_{\hat{n}}[n] \approx \sum_{k=1}^{p} \bar{\boldsymbol{\alpha}}_{\hat{m}}(k)s_{\hat{n}}[n-k] + e_{\hat{n}}[n] \tag{12}$$

4. RESULTS

An example of vibrated speech before and after processing is shown in Figure 3. The restored speech is free of amplitude, frequency and formant modulations, and is perceptually clearer. To test the results numerically we ran instances of single vowels through a classifier. We used 10 vowel classes with 6 ms time frame MFCC's as feature vectors. For each speaker we trained an SVM on the other 5 speaker's clean vowels (∼30,000 time frames total), and tested on the target speaker's clean(∼6,000 time frames), vibrated, and corresponding de-vibrated vowels (∼25,000 time frames). Note that this is a much smaller number of speakers than should normally be used for this problem, however this method was used simply as a proof of concept. The results are shown in Table 4. The average clean accuracy is low due to the small amount of variation in the test data but the overall trend is still present. There is a significant drop in classification accuracy from clean to vibrated, and a consistent improvement from vibrated to de-vibrated. While this is by no means an complete investigation of the effects of vibration on ASR accuracy, these results indicate that the proposed method does not hurt and may improve accuracy.

This method was also tested on data that was vibrated synthetically based upon the model proposed. The parameters are consistently estimated accurately within a small tolerance level, and the original clean speech is restored.

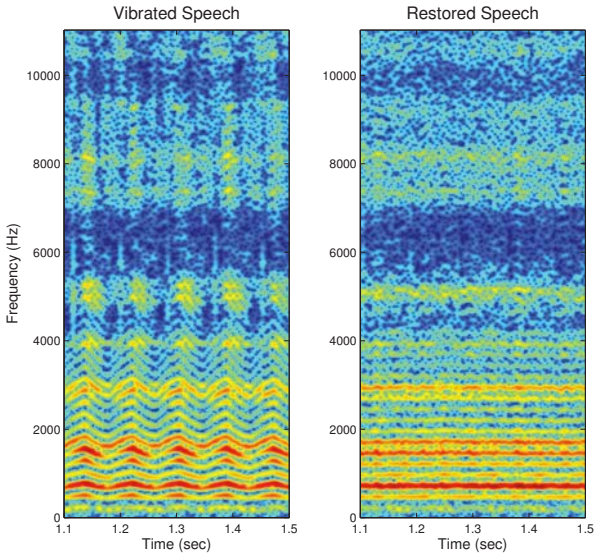

Fig. 3. Spectrograms of the vibrated phoneme [ʌ] at 12 Hz, $0.7g$ (left) and restored phoneme (right).

While a limitation of this method is that it requires a full vibration period, it does not pose a problem in practice. Phonemes sustained for less than one vibration period do not last long enough to be audibly affected by the vibration. In most cases, this method introduces a small amount of noisiness (similar to additive white noise) due to the final filter coefficient smoothing step. However the overall quality of the speech is better after processing despite the addition of noise. The noise is more apparent for male compared to female speakers.

5. CONCLUSIONS

This paper has presented a model that helps provide a better understanding of the different effects of sinusoidal whole-body vibration on speech signals, and a method to remove the effect that shows promise to improve intelligibility.

This work has focused on the ideal case where the applied vibration is sinusoidal at constant frequency and amplitude. However, in practice the vibration is not always this simple. A natural extension of this work would be to broaden the vibration model and the proposed inversion method to handle different types of vibration, such as complex or random. Given the acceleration of the environment over time (as could be measured in real time with an accelerometer) the same model could apply such that the amplitude, frequency, and formants change proportional to the acceleration.

6. REFERENCES

[1] Michael J Griffin, *Handbook of human vibration*, Academic press, 1990.

[2] ANSI S3.2-1989, *Method for measuring the intelligibility of speech over communication systems.*, American National Standards Institute, R1999.

[3] Durand Begault, "Effect of whole-body vibration on speech. part ii: Effect on intelligibility," in *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

[4] NASA/SP-2010-3407, *NASA Human Integration Design Handbook (HIDH)*, Washington, D.C.: NASA, 2010.

[5] C.W. Nixon, "Influence of selected vibrations on speech i. range of 10 cps to 50 cps.," *The Journal of Auditory Research*, vol. 2, pp. 247–266, 1962.

[6] Charles W Nixon and Henry C Sommer, *Influence of Selected Vibrations upon Speech (Range of 2 cps-20 cps and Random)*, Aerospace Medical Research Laboratories, 1963.

[7] Ch W Nixon and HC Sommer, "Influence of selected vibrations upon speech iii. range of 6 cps to 20 cps for semi-supine talkers," *Aerospace medicine*, vol. 34, pp. 1012–1017, 1963.

[8] Robert J Teare, "Human hearing and speech during whole-body vibration," Tech. Rep., DTIC Document, 1963.

[9] Durand Begault, "Effect of whole-body vibration on speech. part i: Stimuli recording and speech analysis," in *Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

[10] Robert McAulay and Thomas Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744–754, 1986.