

SPEECH ANALYSIS USING MODULATION-BASED FEATURES FOR DETECTING DECEPTION

K. Gopalan¹ and S. Wenndt²

¹Purdue University Calumet, Hammond, IN 46321, U.S.A.

²Air Force Research Laboratory, Rome, NY 13441, U.S.A.

ABSTRACT

This paper reports on the initial results of analysis of speech features for detecting deception from speech utterances. The features are derived from the amplitude- and frequency-modulation (AM and FM) behavior of speech centered at different nominal formant frequencies. Analysis of truthful and deceptive speech, both established a posteriori, by a male speaker under jeopardy shows that Teager energy-related features, amplitude envelope-based parameters, and formant variations and bandwidths at selected frequencies have a potential to show subtle variations indicative of deception. For comparison, trajectory of the fundamental frequency of voicing was found to show relatively minor variations between truthful and deceptive speech.

Index Terms – Deceptive speech, modulation model, Teager energy

1. INTRODUCTION

Detection of deception by a speaker based on his or her voice has many applications including security and job screening. Studies have well established that human speech has emotion and other nonlinguistic information encoded in it. Increased activation of the sympathetic nervous system or the parasympathetic nervous system is observed to occur when a speaker is angry, fearful, sad, etc [1]. This increased activation leads to changes in heart rate and blood pressure, and also to tremor in muscle activity. Consequently, the articulatory and respiratory movements for speech production are affected. Similar psychophysiological changes are generally considered to occur during speech under deception. The spectral characteristics of the resulting “stressed” speech are, in general, observed to have increased fundamental frequency F0, increased amplitude, and decreased speech duration. The extent of variation in F0, however, has been shown to depend on the speaker and the type of stress. Thus, detection of deception based on speech has been a challenging task in the area of computer speech

processing. Other manifestations of stress in speech include variations in the formants and their bandwidths, increase in high frequency energy, and changes in the glottal pulse shape. This paper reports on the preliminary results for detecting deception using features from an AM-FM model of speech.

2. MODULATION MODEL OF SPEECH

Representation of a speech frame as a combination of amplitude and frequency modulated signal has been used for analyzing speech and speaker recognition [2, 4] as well as for analyzing and classifying psychophysiological stress in a speaker [2 - 5].

In this model, speech frames are represented as a linear combination of amplitude and frequency modulated signals around each formant F_k as given by

$$x(t) = \sum_{k=1}^N a_k(t) \cos(\omega_k t + \Phi_k(t)) \quad (1)$$

where N represents the number of formants in the speech signal $x(t)$. Amplitude modulation (AM) is given by the time-varying amplitude or envelope $a_k(t)$ at the k^{th} formant frequency ω_k which is modulated by the frequency modulation (FM) term corresponding to the total instantaneous frequency of $\omega_i = \omega_k + \frac{d\Phi_k}{dt}$.

Both the instantaneous frequency (IF) and the amplitude envelope (AE) around a formant are derived from Kaiser’s Teager energy operator $\Psi()$ given by

$$\psi[s(n)] = s^2(n) - s(n-1)s(n+1) \quad (2)$$

From the energies of a speech signal frame bandpass-filtered around a formant using a narrow band filter, and its time-shifted versions, the instantaneous frequency Ω_i and the amplitude envelope $|a(n)|$ are obtained using the energy separation algorithm [2, 4] as

$$\Omega_i(n) \approx \arcsin \sqrt{\frac{\psi[s(n+1)] - \psi[s(n-1)]}{4\psi[s(n)]}} \quad \text{rad/s} \quad (3)$$

$$|a(n)| \approx \frac{2\psi[s(n)]}{\sqrt{\psi[s(n+1)] - \psi[s(n-1)]}} \quad (4)$$

Studies of the IF and AE showed that features based on the above AM-FM demodulation model may indicate variations in the emotional stress of the speaker [3]. In particular, the spectrum of AE showed higher peak frequencies at higher levels of stress; additionally, the spectra of both IF and AE, in general, followed the fundamental frequency F0, with generally increased values for F0, formants and formant bandwidths at elevated heart rates.

Inasmuch as lying is generally considered to increase the stress level of the speaker, similar variations in modulation parameters were expected for deceptive speech.

3. DATABASE

The database used was obtained from the audio recording of a male suspect under criminal investigation. The suspect was determined to have given deceptive statements under questioning during polygraph testing. Audio recordings of two sessions of polygraph testing with the same questions by the investigator and the same responses by the suspect were available. For analysis, two pairs of truthful utterances, or ground truth (labeled R4 and R5), and deceptive utterances (labeled R7 and R9), of the word “no” from each recording were selected. These utterances were sampled at the rate of 16,000 samples/s.

4. MODULATION ANALYSIS OF TRUTHFUL AND DECEPTIVE SPEECH

Modulation analyses were performed on the truthful and deceptive utterances at three formant frequencies. Since the utterance ‘No’ is a nasal–vowel combination, formants for the nasalized ‘oh’ were estimated from short-time Fourier spectra of several frames. The following values were estimated for the fundamental frequency and the formants: $F_0 \approx 110$ Hz, $F_1 \approx 500$ Hz, $F_2 \approx 800$ Hz, and $F_3 \approx 1100$ Hz. The formant around 800 Hz, however, was not noticeable for all the utterances. Figs. 1 and 2 show the formants observed using short time spectral analysis.

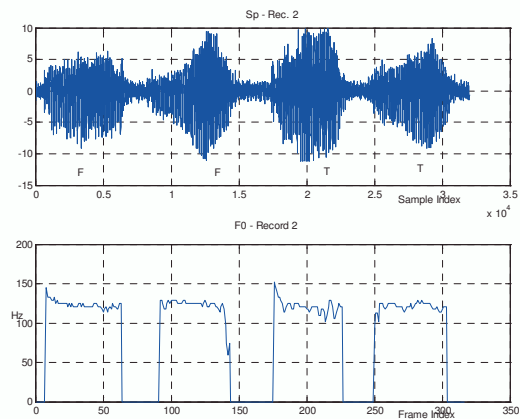


Fig. 1. Fundamental frequency F0 for four concatenated utterances of ‘No’ recorded in a single polygraph session. The first two utterances in the record were deceptive and the last two were truthful.

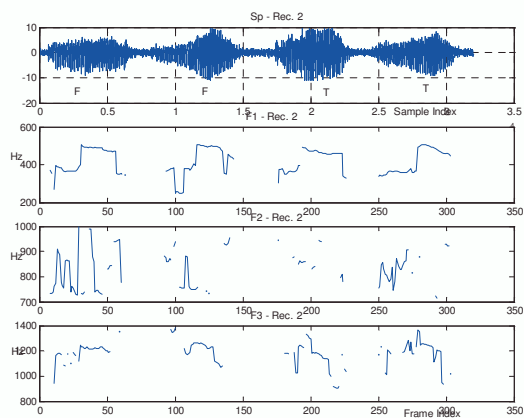


Fig. 2. First three formants for the four concatenated utterances of ‘No’.

The relatively constant F0, which is generally considered to provide an indication of emotional [5] and psychophysiological [6, 7] stress, illustrates the difficulty of detecting deception using F0 for the speaker under consideration.

For modulation analysis, frames of 31.25 ms (500 samples) were analyzed every 6.25 ms (100 samples) after bandpass-filtering around a formant. The following

features were obtained for each frame of bandpass-filtered speech: amplitude modulation index, peak Teager energy, maximum frequency of variation of Teager energy, weighted formant, weighted formant bandwidth, prediction error from a 10th order linear prediction model of AE, and weighted envelope frequency.

The weighted formant F_w is defined by:

$$F_w = \frac{\overline{f_i [a(t)]^2}}{[a(t)]^2} = \frac{\left(\frac{1}{T}\right) \int_{t_0}^{t_0+T} f_i(t) [a(t)]^2 dt}{msq(a)}, \quad (5)$$

where f_i = instantaneous frequency and a = amplitude envelope.

The weighted formant bandwidth B_{wf} is defined by

$$[B_{wf}]^2 = \frac{[f_i - F_w]^2 [a(t)]^2}{[a(t)]^2} = \frac{\left(\frac{1}{T}\right) \int_{t_0}^{t_0+T} [f_i - F_w]^2 [a(t)]^2 dt}{msq(a)} \quad (6)$$

The weighted envelope frequency ω_w is given by

$$\omega_w = \frac{\int_{-\infty}^{\infty} \omega |A(\omega)| d\omega}{\int_{-\infty}^{\infty} \omega d\omega}, \quad (7)$$

where $A(\omega)$ = spectrum of A .

5. RESULTS OBSERVED

Demodulation around the first formant of 400 Hz appeared to show a slight increase in the weighted formant for the truthful utterances relative to the deceptive ones, in general. Not all other features, however, showed a distinction between the pairs of utterances for the speaker. While a definitive conclusion could not be reached without additional data sets, results of modulation analysis around the first formant, shown in Fig. 3, appeared to warrant further study. Initially, weighted formant bandwidth B_{wf} shown in Fig. 4 for one of the two records, for example, did not indicate much dissimilarity between ground truth and deception; on closer inspection of the speech portion excluding silence at the beginning and end of each utterance, however, a discernible difference in the statistics of B_{wf} can be observed. Table 1, which lists its mean, median and standard deviation, shows a slightly increased mean and standard deviation for deceptive speech.

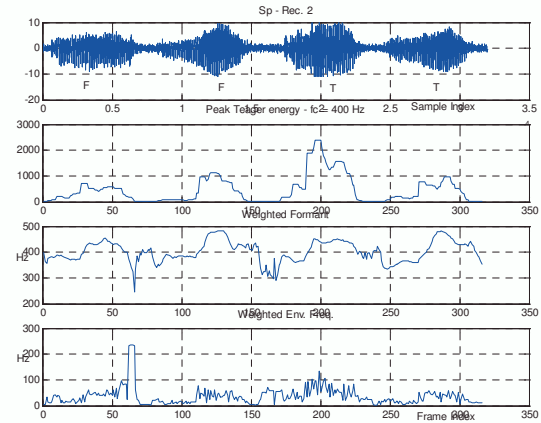


Fig. 3. Speech (top), peak Teager energy, weighted formant and weighted envelope frequency (bottom) around F1 \approx 400 Hz. The first two utterances in the record were deceptive and the last two were truthful.

Table 1

Statistics of Weighted Formant bandwidth for two records around F1 \approx 400 Hz

Record 1

	R4 (F)	R5 (F)	R7 (T)	R9 (T)
Mean, Hz	26.92	24.96	21.02	21.91
Median, Hz	23.41	23.74	20.52	23.48
Standard Deviation, Hz	18.2	10.48	8.36	7.29

Record 2

	R4 (F)	R5 (F)	R7 (T)	R9 (T)
Mean, Hz	35.63	23.45	35.44	24.07
Median, Hz	21.96	23.75	27.32	23.40
Standard Deviation, Hz	36.31	9.66	23.12	5.54

The speaker's responses of 'No' to jeopardy-attached questions during two polygraph recording sessions are listed as R4, R5, etc. with R4 and R5 independently verified as false (F) and R7 and R9 as truthful (T) utterances.

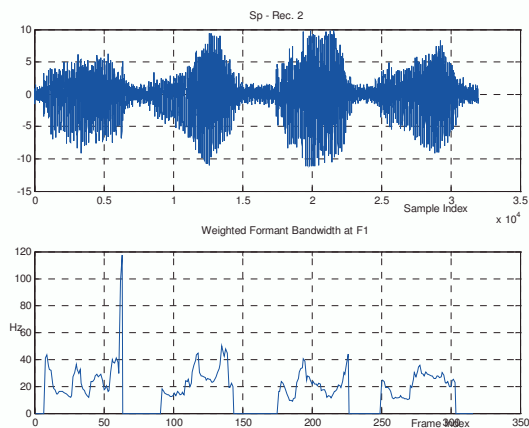


Fig. 4. Weighted formant bandwidth around $F1 \approx 400$ Hz.

The second formant $F2$ of approximately 700 Hz was not present in all voiced frames of all four utterances of a record, as can be seen in Fig. 2. While this in itself may be sufficient to characterize a deceptive utterance compared to a truthful utterance, more data sets are needed to confirm the presence or absence of $F2$ in deceptive and truthful utterances. Demodulation around $F2$ did not show any distinguishing feature for ground truth and deception-indicated speech.

Analysis around the third formant of approximately 1100 Hz showed higher Teager energy profiles for truthful statements of the one of the two records; the profiles, however, were not consistent for the other record. Similar variations were observed for the weighted formant and the weighted envelope frequency. Weighted formant bandwidth B_{wfs} generally considered to be indicative of stress, did not show any consistent pattern between truthful and deceptive utterances.

6. CONCLUSION

Based on the preliminary analysis conducted on the speech from a single speaker, it appears that parameters based on Teager energy and AM-FM demodulation have a potential to discriminate deceptive speech from a

truthful utterance. Inasmuch as pitch tracking is generally considered to correlate with stress (as well as in Lombard speech), the low variability of $F0$ for the database used indicates the difficulty of relying on $F0$ alone for detecting deception. Variations, albeit small, in the modulation-based features, such as the weighted formant bandwidth, for example, indicate the possibility of discriminating between truthful and deceptive utterances. It is expected that a fusion of AM-FM features with short-time spectral and temporal features will contribute to the goal of detecting deception in speech.

Acknowledgement: The authors gratefully acknowledge the help provided by Clifton Hopkins, Law Enforcement Analysis Facility, Lockheed Martin, Rome, NY.

7. REFERENCES

- [1] C.E. Williams, and K.N. Stevens, "Vocal Correlates of Emotional Stress," in *Speech Evaluation Psychiatry*, J.K Darby, Jr. (Ed.), Grune & Stratton, Inc., 1981.
- [2] P. Maragos, J.F. Kaiser and T.F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3025-3051, 1993.
- [3] K. Gopalan, "Amplitude and Frequency Modulation Characteristics of Stressed Speech," Final Report, AFOSR Summer Faculty Research Program, Bolling AFB, July 1998.
- [4] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, No. 3, pp. 201 - 216, 2001.
- [5] K. Gopalan, "On the Effect of Stress on Certain Modulation Parameters of Speech," *Proc. of the 26th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, Salt Lake City, UT, May 2001.
- [6] A. Protopapas, and P. Lieberman, "Effects of Vocal $F0$ Manipulations on Perceived Emotional Stress," *Technical Proceedings, Workshop on Speech under Stress Conditions*, NATO Defence Research Group, pp. 23-1 - 23-4, 1995.
- [7] P. Benson, "Analysis of the Acoustic Correlates of Stress from an Operational Aviation Emergency," *ibid*, pp. 25-1 - 25-4.