An addendum to "Effects of noise on speech production: Acoustic and perceptual analyses" [J. Acoust. Soc. Am. 84, 917–928 (1988)]

W. Van Summers, Keith Johnson, David B. Pisoni, and Robert H. Bernacki
Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405

(Received 16 June 1989; accepted for publication 6 July 1989)

The authors respond to Fitch's comments [H. Fitch, J. Acoust. Soc. Am. **86**, 2017–2019 (1989)] on an earlier paper. New analyses are presented to address the question of whether $F1$ differences observed in the original report are an artifact of linear predictive coding (LPC) analysis techniques. Contrary to Fitch's claims, the results suggest that the $F1$ differences originally reported are, in fact, due to changes in vocal tract resonance characteristics. It is concluded that there are important acoustic-phonetic differences in speech when talkers speak in noise. These differences reflect changes in both glottal and supraglottal events that are designed to maintain speech intelligibility under adverse conditions.

PACS numbers: 43.70.Fq, 43.70.Gr

INTRODUCTION

This issue of the *Journal* contains a Letter of the Editor by Fitch (1989) that is critical of several aspects of our recent report dealing with the acoustic characteristics and intelligibility of speech produced in noise (Summers *et al.*, 1988). Fitch's primary concerns are as follows: First, our discussion should have focused more on the underlying physiological bases of the acoustic differences we reported; in particular, there was little mention of how our data on spectral tilt relate to previous studies in which changes in tilt are associated with changes in vocal effort and with specific changes in glottal waveform shape; second, we did not adequately discuss the possible relationship between our spectral tilt data and our intelligibility results; and, third, our results concerning $F1$ frequency changes in the presence of noise may have been an artifact of the LPC methods used to estimate formant frequencies and may not reflect true changes in vocal tract resonances. We address each of these criticisms, in turn, below.

Fitch's initial criticism of our paper is that our discussion of differences in the acoustic properties of Lombard versus normal speech did not address the underlying source of these differences in speech production. Because we did not collect physiological data, it was not possible to specify exactly what articulatory changes might have been made by our talkers. We agree that, in order to understand the articulatory differences between Lombard and normal speech, it is necessary to gather both articulatory and acoustic data. However, we believe that our findings represent an important contribution to understanding acoustic and perceptual consequences of speaking in noise because they show that, in addition to previously reported acoustic properties of Lombard speech (increased amplitude, $F0$, and duration), there are also reliable spectral differences between Lombard and normal speech.

According to Fitch, the changes in spectral tilt and $F1$ frequencies that we report may all be related to changes in laryngeal behavior. Concerning changes in spectral tilt, we conclude that Fitch is basically correct in her criticisms of our work, although the issue is not as simple as she suggests. Concerning her criticisms of the $F1$ differences, we disagree more strongly. We will argue that the different average $F1$ values which we found for Lombard speech versus speech in the clear reflect real changes in supralaryngeal articulation and are not due to possible artifacts of the measurement procedures. Although we do not have articulatory data to support this claim, we are able to reject Fitch's explanation of our $F1$ data.

I. SPECTRAL TILT AND VOCAL EFFORT

Fitch begins by pointing out an important relationship between our data on spectral tilt and previous research on vocal effort. She compares our results with those of Fant (1959, 1960) and others who reported differences in spectral tilt associated with changes in effort. The changes in spectral tilt reported by Fant are similar to those reported in our paper for Lombard speech. Fant states that these differences in tilt are due to changes in the abruptness of the transitions from the closed to open (and open to closed) phase of the glottal cycle. A recent study of glottal waveforms for speech produced in a variety of environments agrees with this description for loud versus normal speech (Cummings *et al.*, 1989). However, Cummings *et al.* did not find comparable differences between Lombard speech and normal speech. Indeed, they reported that glottal waveforms for Lombard speech are more similar to those of "clear speech" than "loud speech." These results suggest that speech produced under explicit instructions to increase vocal effort may differ in a number of ways from Lombard speech. In Lombard speech, speakers apparently modify the quality of their speech to maintain intelligibility and do not simply increase vocal effort (see Lane and Tranel, 1971; Lane *et al.*, 1970). Therefore, the articulatory changes that occur as a response to increased noise in the environment may not be generalizable to all situations in which vocal effort is increased. Nevertheless, the similarity between our spectral tilt data and Fant's results is important and should have been discussed in our original article.

A related point concerns how changes in spectral tilt may influence intelligibility. Fitch cites an argument originally made by Rostolland (1982) that the decrease in spectral tilt observed in loud speech may improve intelligibility because the frequency region receiving the largest amplitude boost is the region in which hearing sensitivity is greatest. Fitch suggests that spectral tilt differences may explain why Lombard speech was more intelligible in our perceptual experiments. Once again, Fitch's point is an important one that should have been mentioned. However, while decreases in spectral tilt may be beneficial to intelligibility, it is also very likely that durational increases observed in Lombard speech also contribute substantially to increased intelligibility. These durational effects appear to be analogous to those found in "clear speech" (Chen, 1980; Picheny *et al.*, 1986), where subjects are specifically instructed to produce highly intelligible utterances. In short, while changes in spectral tilt may be one factor producing increased intelligibility for Lombard speech, it seems unlikely that these changes are the only ones relevant.

II. *F*0 AND LPC ESTIMATES OF *F*1

In the second part of her letter, Fitch focuses on our analyses of formant frequencies, particularly the first formant. Using LPC techniques to estimate formant frequencies, we observed increases in $F1$ for utterances produced in noise. Fitch suggests that these differences may be due to the use of LPC modeling and may not actually reflect changes in vocal tract resonances. She points out two types of biases inherent in LPC modeling that may have played a role. First, LPC pole values (that are used in estimating formant values) tend to have a bias towards the nearest harmonic of the fundamental. As a result, estimates of formant frequency for a given utterance are influenced by its fundamental frequency.[1] The effect of $F0$ on formant estimates is most pronounced for high $F0$ values because of the wider spacing of harmonics. For this reason, the bias was of less concern in our study than if we had examined women's or children's speech. Nevertheless, Fitch points out that our data do show increases in fundamental frequency for utterances produced in noise. She argues further that the increases in $F0$ may have caused the LPC-based estimates of $F1$ to increase accordingly without any actual change in vocal tract resonant frequencies.

Fitch goes on to describe the LPC estimate of $F1$ as "riding up with the harmonics." She seems to be suggesting here that increases in $F0$ are consistently related to increases in formant peak estimates. However, an increase in $F0$ does not necessarily cause harmonics near a given pole to also increase. Consider a vocal tract configuration with an $F1$ frequency at 500 Hz. If $F0$ is 100 Hz, there will be a harmonic that coincides precisely with the formant peak. The LPC model will presumably do an accurate job of estimating $F1$ frequency in this case. However, if $F0 = 120$, the nearest harmonic will be at 480 Hz and the LPC estimate of $F1$ will "migrate" towards this value. Thus, in some circumstances, an increase in $F0$ may actually lead to a decrease in the LCP estimate of $F1$. This observation calls into question the direct link that Fitch attempts to draw between our $F0$ and $F1$ results. While a change in $F0$ may influence estimates of formant frequencies, $F0$ increases will not necessarily bias estimates of formant frequency upwards (see Atal and Schroeder, 1974).

Nevertheless, some of the data reported in our original article did suggest a close relationship between $F0$ and $F1$: Speaker SC showed higher mean values for both $F0$ and $F1$ when speaking in noise, whereas speaker MD showed fairly stable $F0$ and $F1$ values across quiet and noise conditions. We therefore reported an analysis of covariances (ANCOVA) on the data for SC, to examine the relationship between noise condition and $F1$, with $F0$ treated as a covariate. In this analysis, the linear relationship between $F0$ and $F1$ was accounted for prior to examining whether $F1$ frequency varied across noise conditions. The results suggested that $F1$ tended to increase when speaking in noise apart from any increase in $F0$. Fitch finds this attempt to dissociate $F0$ from $F1$ unconvincing. She suggests that the relationship between $F0$ and $F1$ may not be linear since $F1$ values for various utterances are not associated with the same harmonic(s) of $F0$. Her point is that a given increase in $F0$ will cause a greater frequency change in higher harmonics of $F0$ than in lower harmonics and therefore may have a greater effect on $F1$ measurements for vowels with high $F1$ frequencies than for vowels with low $F1$ values (if $F1$ estimates are indeed "riding up with the harmonics"). Fitting a single regression line to the ($F0$, $F1$) data may therefore provide a fairly poor fit when vowels containing various $F1$ frequencies are analyzed together. To test this possibility, we conducted separate analyses each of the ten vocabulary items produced by speaker SC. For each word, an analysis of covariance was conducted to examine the effect of noise condition of $F1$ frequency with $F0$ treated as a covariate. Within each noise condition, overall mean $F1$ values across vocabulary items were then computed based on the adjusted means from these separate analyses. These adjusted $F1$ values are as follows: quiet—490.4 Hz; 80-dB noise—520.5 Hz; 90-dB noise—518.4 Hz; 100-dB noise—532 Hz. These values are similar to the values from the ANCOVA originally reported in Summers *et al.* (1988). Clearly, the tendency for $F1$ to increase in noise is still present.

Fitch claims that LPC-based estimates of $F1$ will increase with increases in $F0$. If this is true, then there should be a positive relationship between $F0$ and $F1$ for tokens within a given condition as well as across conditions. We therefore examined the relationship between $F0$ and $F1$ across the five tokens of each word produced in each noise condition by SC. The $F0$ and $F1$ values for each token were used in a linear regression analysis, with $F0$ used as a predictor of $F1$. Separate analyses were carried out for each vocabulary item within each noise condition for a total of 40 analyses (ten items × four noise conditions). If $F0$ increases are consistently associated with $F1$ increases, the resulting regression lines should have positive slopes. Of the 40 regression lines computed, 22 showed positive slopes, while 18 showed negative slopes. Clearly, the results fail to establish any strong positive relationship between $F0$ and LPC-based estimates of $F1$ for tokens of an utterance produced in a given noise condition.

III. SPECTRAL TILT AND LPC ESTIMATES OF $F1$

The second type of bias that Fitch suggests may have influenced our $F1$ results relates to how changes in spectral tilt influence the LPC model. Changes in spectral tilt, which we observed in our study, produce differences in the relative amplitudes of harmonics near a given formant. This, in turn, could affect the LPC estimates of formant frequency without any change in vocal resonance characteristics. On initial inspection, Fitch's suggestion does not appear to be consistent with our data, since speaker MD showed decreases in spectral tilt without accompanying changes in $F1$ frequency.

To further test the relationship between spectral tilt and $F1$ frequency in our data, we conducted an analysis of covariance on speaker SC's $F1$ data, treating noise condition and utterance as independent variables and spectral tilt[2] as a covariate. The adjusted mean $F1$ frequencies from this analysis were: quiet—499 Hz; 80-dB noise—522.6 Hz; 90-dB noise—514.3 Hz; 100-dB noise—525.4 Hz. These results suggest that a portion of the $F1$ increase observed in the data between the quiet and noise conditions is related to changes in spectral tilt; that is, the change in $F1$ frequency across conditions is reduced when spectral tilt is entered as a covariate in the analysis. Nevertheless, as in the analysis of variance originally reported, the main effect of noise on $F1$ frequency was statistically significant in the analysis of covariance [$F(3,159) = 4.98, p < 0.0025$]. Thus the effects of noise on $F1$ for speaker SC cannot be completely accounted for by variability in spectral tilt.

In discussing the influence of tilt on $F1$ measurements, Fitch compares our results with data reported by Makhoul and Wolf (1972), who examined the effect of preemphasis on formant estimates. Makhoul and Wolf report that preemphasis, which has a substantial effect on spectral tilt in the $F1$ region, also affects $F1$ frequency measurements.[3] If the changes in spectral tilt that we observed when talkers speak in noise are similar in kind and magnitude to changes due to preemphasis, it would follow that the $F1$ changes we observed might then appropriately be ascribed to changes in tilt. However, the change of spectral tilt which occurs by changing the preemphasis factor is not the same as the change of specral tilt that results from changing the glottal source function. Energy levels at low frequencies (around $F1$) are most affected by changes in preemphasis. On the other hand, in naturally occurring changes of the glottal sources spectrum, low frequencies are not affected as much as high frequencies (Rostolland, 1982).

We manipulated the spectral tilt of SC's utterances produced in quiet in order to determine how a change in spectral tilt approximating the change observed in noise would affect $F1$ frequency measurements. A 51-pole filter was designed to model of the difference of the long term spectra of vowels produced in the 100-dB noise condition and vowels produced in the quiet (an average difference of approximately 2 dB per octave). The items produced in the quiet were then filtered and reanalyzed. The slope of the spectrum of these "filtered-quiet" vowels was virtually identical to the slope for SC's vowels produced in 100-dB noise. If the slope of the spectrum has the kind of effect on LPC estimates of $F1$ that Fitch claims, then we would expect the average $F1$ of the filtered vowels to match that of the vowels produced in noise. This was not the case. The vowels produced in noise had an average $F1$ of 531 Hz, while the filtered-quiet vowels had an average $F1$ of 507 Hz.[4]

We then examined how an equivalent (2 dB per octave) change in tilt due to preemphasis might influence spectral shape and estimates of $F1$ frequency. In our original analysis, we used a preemphasis factor of 99% throughout. By reanalyzing the utterances produced in the 100-dB noise condition using a 66% preemphasis factor, we altered the overall spectral tilt of these utterances by approximately 2 dB per octave, giving them an average tilt approximating that of the items produced in quiet. As already mentioned, this method of altering spectral tilt has its main effect on the low-frequency portion of the spectrum. Thus it is not surprising that this manipulation had a pronounced effect on the LPC estimate of $F1$. The average estimated $F1$ for the items produced in noise was 531 Hz when preemphasis was 99%, while the average $F1$ of the same items with preemphasis of 66% was 494 Hz. This effect of preemphasis on estimates of $F1$ replicates Makhoul and Wolf's (1972) findings.[5] Our conclusion, then, is that not all sources of spectral tilt have comparable effects on LPC estimates of $F1$. The change in spectral tilt produced by preemphasis techniques has a much greater effect on $F1$ than the change in spectral tilt produced by adjustment at the glottis.

IV. $F0$ and $F1$ IN PERCEPTION

We have mentioned that the LPC estimate of $F1$ fluctuates around the actual $F1$ as a function of $F0$ (Atal and Schroeder, 1974). It is also important to note that perceived $F1$ varies as a function $F0$ and that this perceptual variation (for the range of $F0$ variation present in our data) is very similar to the variation found in LPC estimates of $F1$. Changes in the relative amplitudes and spacing of harmonics in the $F1$ region influence hearers' estimates of $F1$ (Carlson *et al.*,1975; Darwin and Gardner, 1985; Chistovich, 1985; Sundberg and Gauffin, 1978; Assmann and Nearey, 1987). Darwin and Gardner (1985) found that listeners' estimates of $F1$ (as determined in a matching task) were shifted down when the amplitude of a harmonic somewhat below the actual $F1$ was increased and that estimated $F1$ was shifted up when the amplitude of a harmonic somewhat above the actual $F1$ was increased. So, in a sense, it can be said that listeners' estimates of $F1$ "ride up the harmonics," but, as with LPC estimation, the location of the harmonics relative to the actual vocal tract resonance determines whether the estimated $F1$ will be lower or higher than the actual $F1$. In work that related perceptual estimation of $F1$ and LPC estimation, Assman and Nearey (1987) found that LPC estimates of $F1$ for tokens in which the amplitudes of harmonics in the $F1$ region were manipulated correspond very closely to listeners' estimates of $F1$ (in a matching task). Again, the fluctuation in LPC estimates of $F1$ that have to do with changing $F0$ may have a parallel in the perceptual estimation of $F1$ (Ref. 6).

V. CONCLUSION

In summary, Fitch criticizes our recent article on two counts. First, she identifies several points in our presentation in which our findings should have been related to previously published work on increased vocal effort and shouted speech. We are generally in agreement with these criticisms. The similarity between our spectral tilt data and previous data examining vocal effort should have been noted. While we have indicated certain reservations about the appropriateness of claiming that Lombard speech is the same as other forms of speech produced with increased vocal effort, the similarity between our spectral tilt data and that of Fant (1959, 1960) should have been discussed. Also, we should have discussed Rostolland's (1982) suggestions concerning the possible relationship between spectral tilt and speech intelligibility. Clearly, if Fitch had served as a reviewer of our paper, these suggestions would have been included and no doubt would have improved the discussion.

Fitch's second criticism focuses more on the scientific merit of our study than on the completeness of our discussion of its results. She points out several possible biases inherent in our use of LPC methods of estimating formant frequencies and goes on to suggest that these biases may explain our $F1$ data. As a result, Fitch contends that our findings concerning $F1$ may not reflect real changes in supralaryngeal articulation and are therefore uninteresting. We disagree with both of these criticisms. The additional analyses we report here demonstrate that the $F1$ differences we originally found were not simply due to changes in $F0$ or spectral tilt.

Characterizing the acoustic changes in speech that occur when talkers speak in noise is a research problem that has received relatively little attention in the past (however, see Stanton, 1988). Few investigations have been concerned with the consequences of these acoustics changes for speech perception and speech recognition. While information concerning articulatory behavior can be inferred from acoustic data, more direct physiological studies of speech production in severe environments are clearly needed to learn more about how talkers adjust and modify the way they speak when there are changes in their immediate acoustic environment. Most of our current knowledge about speech production and perception has been obtained with cooperative talkers in benign environments with little if any cognitive load or stress. However, most of real-world speech communication is carried out in less than the idealized conditions typically found in the laboratory. Despite the criticism raised by Fitch, we believe that our findings are valid and they suggest important acoustic-phonetic differences in speech produced in noise. To our knowledge, these findings have not been reported before in the literature in connection with Lombard speech. We believe our results are directly relevant not only to the development of new and more robust speech recognition algorithms but to our continued search for a better understanding of human speech perception as it takes place in a wide variety of environments.

ACKNOWLEDGMENTS

Preparation of this paper was supported, in part, by a contract with the Armstrong Aerospace Medical Research Laboratory, Wright–Patterson AFB, Ohio, and, in part, by an NIH training grant to Indiana University. We thank Ken Stevens for his suggestions on a number of issues.

[1] The tendency for formant estimates to be biased toward nearby harmonics of the fundamental is not limited to estimates based on LPC modeling. The tendency is also present for formant estimates based on spectrograms (Lindblom, 1962).

[2] The measure of spectral tilt was the slope of a regression line fit to the spectrum of the highest amplitude analysis frame of a given vowel.

[3] Fitch quotes a brief passage from Makhoul and Wolf (1972) regarding the influence of preemphasis on formant frequency measurements. It is interesting to note that the sentence immediately following the quoted material states: "However, these shifts are not significant in general and can be disregarded for many applications."

[4] The unfiltered quiet vowels had an average $F1$ of 488.5 Hz, so it appears that this tilt manipulation may have had some influence on estimated $F1$. However, the $F1$ values estimated by root solving are: noise—524.4; quiet—486.3; filtered quiet—490.6 Hz.

[5] The results are essentially the same for $F1$ estimated by peak picking and root solving. We report estimates made by peak picking because this was the method used in Summers *et al.* (1988).

[6] Another type of perceptual link between $F0$ and $F1$ should be mentioned. Much of the research having to do with $F0$ normalization has assumed the existence of some type of speaker normalization process in speech perception (Miller, 1953; Fujisaki and Kawashima, 1968; Slawson, 1968) rather than a single auditory effect in formant peak estimation (Traunmüller, 1981). Johnson (1989) has found that, when speaker identity is controlled independently, the effect of $F0$ on the perceptual evaluation of $F1$ is quite small. He suggests (following Slawson, 1968) that $F0$ plays a role in vowel identification at two levels: (1) at a psychophysical level in estimating formant peaks and (2) at a cognitive level as speaker characteristics are taken into account during vowel identification. Johnson's (1989) findings suggest that the rather large effects of $F0$ on perceived $F1$ which have been reported before are the result of the second level of processing.

Assmann, P. F., and Nearey, T. M. (**1987**). "Perception of front vowels: The role of harmonics in the first formant region," J. Acoust. Soc. Am. **81**, 520–534.

Atal, B. S., and Schroeder, M. R. (**1974**). "Recent advances in predictive coding—Applications to speech synthesis," Speech Communication Seminar, Stockholm.

Carlson, R., Fant, G., and Granström, B. (**1975**). "Two-formant models, pitch and vowel perception," an *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London).

Chen, F. R. (**1980**). "Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level," unpublished master's thesis, MIT, Cambridge, MA.

Chistovich, L. A. (**1985**). "Central auditory processing of peripheral vowel spectra," J. Acoust. Soc. Am. **77**, 798–805.

Cummings, K. E., Clements, M. A., and Hansen, J. H. L. (**1989**). "Estimation and comparison of the glottal source waveform across stress styles using glottal inverse filtering," IEEE Proc., Southeastcon., April, 1989, pp. 776–781.

Darwin, C. J., and Gardner, R. B. (**1985**). "Which harmonics contribute to the estimation of first formant frequency?, " Speech Commun. **4**, 231–235.

Fant, G. (**1959**). "Acoustic analysis and synthesis of speech with applications to Swedish," Ericsson Tech. **15**, 1–106.

Fant. C. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Fitch, H. (**1989**). "Comments on effects of noise on speech production: Acoustic and perceptual analyses" [J. Acoust. Soc. Am. **84**, 917–928 (1988)]," J. Acoust. Soc. Am. **86**, 2017–2019.

Fujisaki, H., and Kawashima, T. (**1968**). "The role of pitch and higher formants in the perception of vowels," IEEE Trans. Audio and Electroacoust. **16**, (1), 73–77.

Johnson, K. (**1989**). "Intonational context and $F0$ normalization," J. Acoust. Soc. Am. (submitted).

Lane, H. L., and Tranel, B. (**1971**). "The Lombard sign and the role of hearing in speech," J. Speech and Hear. Res. **14**, 677–709.

Lane, H. L., Tranel, B., and Sisson, C. (**1970**). "Regulation of voice communication by sensory dynamics," J. Acoust. Soc. Am. **47**, 618–624.

Lindblom, B. (**1962**). "Accuracy and limitations of sonagraph measurements," in Proc. Fourth Int. Congr. Phon. Sci., Helsinki, 188–202.

Makhoul, J. I., and Wolf, J. J. (**1972**). "Linear prediction and the spectral analysis of speech," Bolt Beranek and Newman Inc., Cambridge, MA, NTIS AD-479066, Rep. 2304.

Miller, R. L. (**1953**). "Auditory tests with synthetic vowels," J. Acoust. Soc. Am. **25**, 114–121.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing II: Acoustic characterstics of clear and conversational speech," J. Speech Hear. Res. **29**, 434–446.

Rostolland, D. (**1982**). "Acoustic features of shouted voice," Acustica **50**, 118–125.

Slawson, A. W. (**1968**). "Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency," J. Acoust. Soc. Am. **43**, 87–101.

Staton, B. J. (**1988**). "Robust recognition of loud and Lombard speech in the fighter cockpit environment," unpublished Ph.D. dissertation, Purdue University.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**, 917–928.

Sundberg, J., and Gauffin, J. (**1978**). "Waveform and spectrum of the glottal source," STL-QPSR **2-3**, 35–50.

Traunmuller, H. (**1981**). "Perceptual dimension of openness in vowels," J. Acoust. Soc. Am. **69**, 1465–1475.