Speech production modifications produced in the presence of low-pass and high-pass filtered noise

Youyi Lu[a)]
Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street,
Sheffield S1 4DP, United Kingdom

Martin Cooke
Ikerbasque (Basque Science Foundation) and Language and Speech Laboratory, Facultad de Letras,
Universidad del País Vasco, 01006 Vitoria, Spain

(Received 12 September 2008; revised 18 June 2009; accepted 18 June 2009)

In the presence of noise, do speakers actively shift their spectral energy distribution to regions least affected by the noise? The current study measured speech level, fundamental frequency, first formant frequency, and spectral center of gravity for read speech produced in the presence of low- and high-pass filtered noise. In both filtering conditions, these acoustic parameters increased relative to speech produced in quiet, a response which creates a release from masking for listeners in the low-pass condition but which actually increases masking in the high-pass noise condition. These results suggest that, at least for read speech, speakers do not adopt production strategies in noise which optimize listeners' information reception but that instead the observed shifts could be a passive response which creates a fortuitous masking release in the low-pass noise. Independent variation in parameters such as F0, F1 and spectral center of gravity may be severely constrained by the increase in vocal effort which accompanies Lombard speech.
© *2009 Acoustical Society of America.* [DOI: 10.1121/1.3179668]

PACS number(s): 43.72.Ar, 43.72.Dv, 43.66.Dc [DOS] Pages: 1495–1499

I. INTRODUCTION

Speakers change the way they speak in the presence of noise (Lombard, 1911), causing, amongst others, an increase in speech level and fundamental frequency (F0), a flattening of spectral tilt (i.e., more energy at higher frequencies), and a tendency for an upward shift of F1 frequency. While the scale of changes in acoustic parameters observed in "Lombard" speech appears to be related to the relative level of the masker (Summers *et al.*, 1988; Tartter *et al.*, 1993), noise maskers with differing spectral shapes and temporal fluctuations have led to consistent changes in speech level, F0, and spectral tilt. For example, different studies employed white noise (Pisoni *et al.*, 1985; Summers *et al.*, 1988; Junqua, 1993), pink noise (Bond *et al.*, 1989; Hansen, 1996), traffic noise (Letowski *et al.*, 1993), multitalker babble (Garnier, 2007), and competing talkers (Lu and Cooke, 2008). One interpretation of the consistency with which various types of noise provoke speech production modifications is that the spectro-temporal properties of the noise may play little or no role in the Lombard effect. Under this view, speakers cannot, or do not, engage in active strategies which take into account the effect of noise at the ears of listeners.

However, other studies have raised the possibility that Lombard speech has an active component. Junqua *et al.* (1998) studied the influence of noise spectral tilt on Lombard speech, with a constant masker level of 85 dB sound pressure level (SPL). Speech level and F0 increased relative to a quiet background when talkers spoke with noise in the background in all conditions of spectral tilt, supporting the notion of a passive Lombard component. On the other hand, the size of the increase in speech level varied with noise spectral tilt. Mokbel (1992) recorded speech in the presence of white noise which was presented either low- or high-pass filtered or without filtering, at a fixed level. An increase in speech energy in frequency regions where the noise energy was most concentrated was observed, suggesting a dependency of the Lombard effect on the noise frequency distribution. However, Mokbel's study involved only one single speaker and did not report detailed changes in acoustic parameters, so it is difficult to appreciate the precise pattern as well as the reliability of the results, given that significant speaker-dependency of speech produced in noise has been observed (Summers *et al.*, 1988; Junqua 1993). However, the studies of Junqua *et al.* (1998) and Mokbel (1992) raise the intriguing possibility that the Lombard effect may have an active component which depends on the spectral characteristics of the background noise. In other words, talkers might use information gained by listening-while-talking to affect purposeful modifications to their speech, perhaps with the goal of improving intelligibility at the ears of the interlocutor.

Lu and Cooke (2008) investigated the effect of N-talker babble noise on speech production for N ranging from 1 (a single competing talker) to "infinity" (speech-shaped stationary noise), and taking in various multitalker babble conditions for intermediate values of N. Consistent with other Lombard studies, an overall shift in the center of gravity (CoG) of energy from lower to higher frequencies was observed at all values of N. Further, listeners found Lombard

[a)] Author to whom correspondence should be addressed. Electronic mail: y.lu@dcs.shef.ac.uk

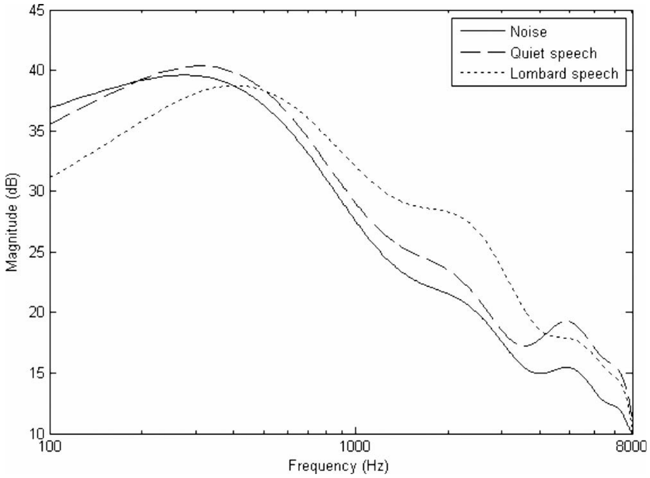

FIG. 1. Long-term average spectra of speech-shaped noise, and speech produced in quiet and noise in Lu and Cooke (2008). Note that the signals have normalized rms energy. A clear Lombard effect of energy shift to higher frequencies relative to quiet speech is visible.

speech substantially more intelligible than speech produced in quiet when both were presented in speech-shaped noise at the same signal-to-noise ratio. Since the long term spectrum of the noise was speech-shaped (for all *N*), an upward shift in CoG causes a degree of release from energetic masking (Fig. 1). Thus, the improvement in intelligibility could be fortuitous, since noise-induced speech changes may coincidentally be in the right direction to be advantageous for the speech-shaped noise maskers. An alternative possibility is that the observed shifts were caused by speakers making an active attempt to place spectral information in locations where it was less likely to be masked. The purpose of the current study was to distinguish these two possibilities.

In the present study, changes in speech production were measured in conditions of low-pass, high-pass, and full-band speech-shaped noise, relative to quiet. If speakers adopt an optimal strategy in order to minimize the effect of noise on listeners, they would be expected to shift their spectral CoG downwards for high-pass filtered noise condition compared to quiet, and in the opposite direction for low-pass noise condition. For each of the high- and low-pass conditions, two noise bandwidths were used to investigate the effect of varying the size of the noise-free part of the spectrum. Again, a "listener-optimal" speaking strategy should lead to greater changes for the smaller noise-free regions because the shift in speech spectral energy would need to be larger to reach the clean parts of the spectrum.

II. SPEECH CORPUS COLLECTION

A. Speech material

Speakers produced sentences defined by the Grid structure used in previous collections of normal (Cooke *et al.*, 2006) and Lombard speech (Lu and Cooke, 2008). Grid specifies simple six-word sentences such as "bin green at K 4 now" or "place red by E 7 please." While Grid sentences are not representative of natural tasks, they control for differences in speaking style and syntax, and the existence of many keyword repetitions allows for cross-condition comparisons of acoustic properties. Talkers produced an identical set of 30 Grid sentences in each of the conditions (see Sec. II B). To introduce some variation and remove any sentence dependency effect, each talker used a different sentence set.

B. Noise backgrounds

Speech was collected in quiet and in the presence of five noise backgrounds, one full-band, two high-pass filtered, and two low-pass filtered. The full-band noise had a spectrum equal to the long-term spectrum of utterances drawn from the 16 female and 18 male talkers of the Grid corpus (Cooke *et al.*, 2006), shown in Fig. 1. Low- and high-pass noise were derived from full-band noise using Chebyshev filter implementations with 0 dB pass-band gain and 60 dB stop-band attenuation, with frequency responses illustrated in Fig. 2. To investigate the effect of the size of the stop-band on speech production in noise, narrow- and wide-band versions of both high- and low-pass noise were generated using cutoff frequencies of 1 and 2 kHz. Note that in the low-pass conditions, the 1 kHz cutoff results in a narrow-band noise while in the high-pass condition the same cutoff leads to a wide-

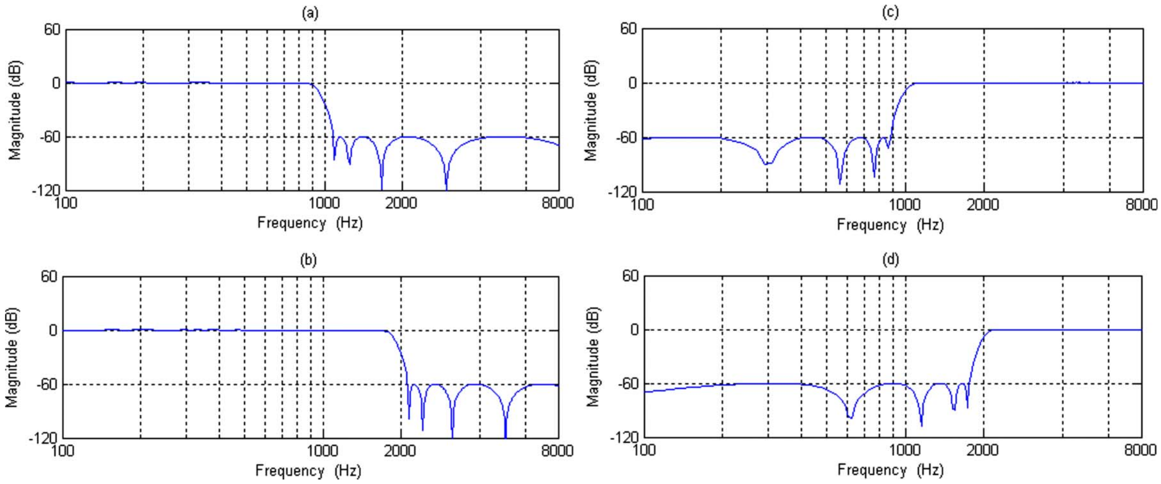

FIG. 2. (Color online) Frequency responses of the low- and high-pass digital filters. Panel (a) and (b) represent the low-pass filters with cutoff frequencies of 1 and 2 kHz respectively. Panel (c) and (d) represent the high-pass filters with cutoff frequencies of 1 and 2 kHz, respectively.

band noise, and vice versa for the 2 kHz cutoff. All maskers were normalized to 89 dB SPL prior to presentation, as measured with a Bruel & Kjaer (B & K) type 2603 sound level meter and B & K type 4153 artificial ear.

C. Talkers

Eight native speakers of British English (four males and four females) drawn from staff and students in the Department of Computer Science at the University of Sheffield participated in the corpus collection. All received a hearing test using a calibrated software audiometer which was used to test each ear separately at the six frequencies: 250, 500, 1000, 2000, 4000, and 8000 Hz. All participants had normal hearing (better than 20 dB hearing level in the range of 250–8000 Hz). Ages ranged from 24 to 48 years (mean: 29.8 years). Ethics permission was obtained following the University of Sheffield Ethics Procedure. Talkers were paid for their participation.

D. Procedure

Corpus collection sessions took place in an industrial acoustics company single-walled acoustically-isolated booth. Speech material was collected using a B & K type 4190 $\frac{1}{2}$ in. microphone coupled with a preamplifier (B & K type 2669) placed 30 cm in front of the talker. The signal was further processed by a conditioning amplifier (B & K Nexus model 2690) prior to digitisation at 25 kHz with a Tucker-Davis Technologies (TDT) RP2.1 system. Simultaneously, maskers were presented diotically over Sennheiser HD 250 Linear II headphones using the TDT system. Talkers wore the headphones throughout, including for the quiet condition. In order to compensate for sound attenuation introduced by the closed ear headphones, the talkers' own voice was fed back via the TDT system and mixed with the noise signal prior to presentation over the headphones. At the beginning of the recording session, each talker was asked to speak freely into the microphone while wearing the headphones. The level of voice feedback was manually adjusted until the talker felt that the overall loudness level matched that when not wearing headphones. Voice feedback level was then held constant for all the recording conditions and talkers were unable to adjust the level.

Sentence collection and masker presentation was under computer control. Talkers were asked to read out sentences presented on a computer screen and had 3 s to produce each sentence. They were allowed to repeat the sentence if they felt it necessary, with the final repetition used for further analysis. In practice, talkers made only a few repetitions in any single condition with maximum of 4 out of 30 sentences and a mean of less than 2. Across-talker means of repetition in the six conditions were not statistically different $[F(1,7)=0.86, p=0.44]$. Maskers were gated with the 3 s recording time. Condition and sentence orders within each condition were randomized. Talkers recorded all the six conditions (i.e., five noise conditions plus quiet) in one session of approximately 20 min.

E. Postprocessing

In order to identify and remove leading and trailing silent intervals of the collected sentences, a set of speaker-independent phoneme-level hidden Markov models was built from speech material in the Grid corpus using the HTK toolkit (Young *et al.*, 1999). These models were used to produce phoneme-level transcriptions of the collected utterances via forced alignment using the HVITE tool in HTK. The leading and trailing silent intervals identified via the alignment process were removed. Transcriptions of the leading and trailing silent intervals for all the utterances were manually inspected and found to be accurate within approximately 15 ms relative to human judgements.

III. ACOUSTIC MEASUREMENTS AND STATISTICAL ANALYSIS

Four acoustic properties were estimated for each utterance. Root mean square (rms) energy, mean fundamental frequency (F0), spectral CoG, and mean first formant (F1) frequency were computed via PRAAT 4.3.24 (Boersma and Weenink, 2005). F0 estimates were provided at 10 ms intervals using an autocorrelation-based method (Boersma, 1993) implemented in the PRAAT program. Spectral CoG was computed on the spectrum of an entire utterance by averaging the frequency spectrum weighted by its power magnitude. Mean F1 frequency was obtained by averaging F1 values estimated for voiced frames using the BURG algorithm (Burg, 1975) implemented in PRAAT. These parameters were selected since reliable changes in these properties have been reported in earlier Lombard studies, and, apart from rms energy, all these properties cue the location of spectral information, which allows the pattern of shifts in spectral energy distribution to be determined.

Across-talker means in quiet, speech-shaped noise and filtered noise conditions for each of the acoustic parameters are shown in Fig. 3. For all parameters and in both low- and high-pass conditions, noise resulted in increases in all parameters. In the low-pass case, little difference between the two filtered and full-band noises is visible, while for high-pass noise, filtered noise tended to result in smaller increases than in the full-band condition. While some variability among the individual talkers was present, similar patterns in each of the acoustic parameters and across backgrounds were observed.

Due to the likelihood of moderate correlations between acoustic parameters such as speech level and both F0 and F1 frequency (Alku *et al.*, 2002; Garnier 2007), multivariate analysis of variance (MANOVA) was used to examine the effect of noise background. Separate MANOVAs were computed for the low- and high-pass cases, with rms energy, F0, F1, and CoG as dependent variables. Initially, MANOVAs with one within-subject factor representing four types of background (quiet, narrow, wide, full) and one between-subject factor (gender) revealed that while gender differences were observed for F0 and F1, the pattern of results was the same for the male and female talkers since no significant interaction was found between gender and background type

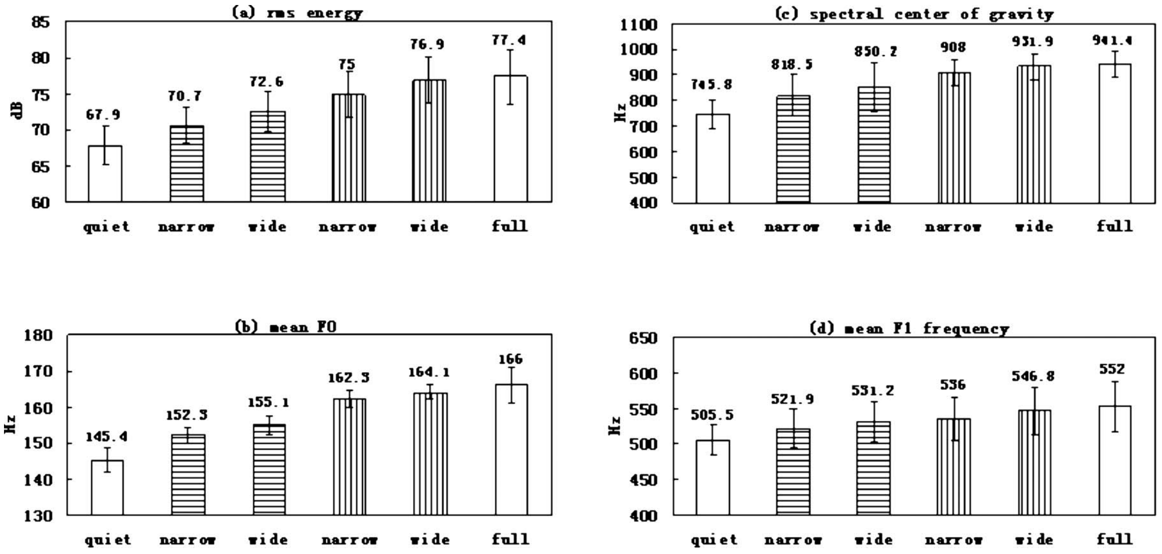

FIG. 3. Acoustic parameter values for quiet, two high-pass noise conditions (shaded bars with horizontal lines) with cutoff frequencies at 2 kHz ("narrow" bandwidth) and 1 kHz ("wide" bandwidth), two low-pass noise conditions (shaded bars with vertical lines) with cutoff frequencies at 1 kHz (narrow bandwidth) and 2 kHz (wide bandwidth), and speech-shaped noise condition ("full" bandwidth). Values shown are means over talkers and error bars indicate 95% confidence intervals.

($p > 0.05$). In order to increase statistical power with the limited number of speakers used in the current study, data for male and female talkers were combined.

For the low-pass case, there was a significant multivariate effect of differences between the four backgrounds {quiet, two low-pass noise, speech-shaped noise} [$F(12,47.9) = 9.37$, $p < 0.001$, $\eta^2 = 0.66$], as well as for the four parameters individually [$F(1.23,8.62) = 49.15$, $p < 0.001$, $\eta^2 = 0.88$ for rms energy; $F(1.38,9.65) = 27.66$, $p < 0.001$, $\eta^2 = 0.80$ for mean F0; $F(1.24,8.67) = 21.87$, $p < 0.01$, $\eta^2 = 0.76$ for CoG; $F(2.05,14.37) = 97.64$, $p < 0.001$, $\eta^2 = 0.93$ for mean F1 frequency]. *Post hoc* pairwise comparisons (here and elsewhere by paired *t*-tests with Bonferroni-adjustment) showed that the quiet condition was significantly different from the rest ($p < 0.01$) for all four parameters. None of the differences between the three noise conditions was statistically significant.

As expected, given the difference between the quiet and full-band conditions, for the high-pass case, the multivariate effect of background type {quiet, two high-pass noise, speech-shaped noise} was also significant [$F(12,47.9) = 5.99$, $p < 0.001$, $\eta^2 = 0.55$]. Of more interest is the confirmation by *post hoc* pairwise comparisons that the high-pass conditions resulted in significant increases in all parameters relative to quiet ($p < 0.05$), and, unlike in the low-pass case, increases were significantly smaller than the full-band condition ($p < 0.05$) apart from the wide-band/full-band comparison for F1 ($p = 0.06$). The tendency, visible in Fig. 3, for the wide-band high-pass noise to provoke larger parameter excursions than the narrow-band high-pass condition was not statistically significant except in the case of rms energy ($p < 0.05$).

IV. DISCUSSION

The current study extends to both low- and high-pass filtered noise backgrounds the finding that talkers modify their productions when exposed to full-band noise. The low-pass conditions resulted in increases in F0 and F1 frequencies, and spectral CoG. While these results are consistent with the hypothesis that speakers were actively avoiding the presence of noise whose spectrum was concentrated at low frequencies, two findings suggest otherwise. First, the full-band and low-pass filtered noise provoked statistically-identical increases in these parameters. One might expect to see a larger amount of shift in the low-pass condition to take advantage of the noise-free part of the spectrum relative to the full-band case. Second, there was no difference between the narrow- and wide-band low-pass conditions, where an active strategy would predict larger increases in the presence of wide-band low-pass noise in order to place spectral energy in the noise-free region.

High-pass filtering conditions also led to clear increases in F0, F1 and spectral CoG, suggesting that speakers are unable to adopt the speaking strategy of adapting speech production to place information-bearing elements of speech in regions devoid of noise. Further, speakers reacted similarly to the wide- and narrow-band conditions, where optimality would suggest that a smaller noise-free spectral region would lead to differential shifts in acoustic parameters. The absence of the "optimal" response to high-pass noise may be attributed to articulatory side-effects of an increase in vocal effort, which was observed in all noise backgrounds. For example, the rise in subglottal pressure needed to increase vocal output leads to an increase in F0 (Schulman, 1985; Gramming *et al.* 1988), and the wider jaw opening in order to increase sound amplitude induces an increase in F1 frequency (Stevens, 2000; Huber and Chandrasekaran, 2006). Thus, the scope for active control of F0 and F1 frequencies might be limited by the stronger desire to increase output level in response to noise.

One surprising aspect of the current study is the fact that noise bandlimited to the region below 1 kHz produced an equivalent Lombard effect as full-band noise. This might result from the upward spread of masking into higher frequencies produced by the 1 kHz low-pass noise, a phenomenon first reported by Egan and Hake (1950). In addition, since all noises employed were presented at the same level, the little difference of Lombard effect between the low-pass filtered and full-band noise conditions appears to support the studies cited in the Introduction which argued that noise level is the dominant component of the Lombard effect. However, the high-pass filtered noise conditions led to a significantly smaller increase in parameters such as rms energy (2.8 and 4.7 dB compared to 7.1 and 9 dB in the low-pass conditions, a difference which probably also accounts for the lower scale of increases in other acoustic parameters given the articulatory constraints discussed above), suggesting that noise level is not the only factor in the Lombard effect. It is possible that the difference in response to high- and low-pass noise reflects the relative importance that these frequency regions have in speech perception or in own-voice monitoring. F0 information is more clearly masked in the low-pass conditions, for instance.

Overall, these findings do not support the idea of an active response to noise. However, there are several aspects of the current task which may have limited the scope or motivation on the part of talkers to exploit noise-free spectral regions. First, noise was gated on and off to coincide with the 3 s recording period. It is possible that speakers were not exposed to noise for long enough to learn about the potential benefit of re-allocating spectral energy. Second, the task for talkers did not involve communication of information, so the notion that talkers were motivated to make things easier for a listener is suspected. Further studies involving communicative tasks and continuous noise backgrounds may lead to different results. Finally, the observed change in speech level produced by noise may act to mask the effect of noise on other parameters. Experiments designed to inhibit the change in vocal effort (e.g., Pick *et al.*, 1989) may provide a more sensitive measure of differential response to the spectral content of the background.

V. CONCLUSION

An effective speaking strategy for the maintenance of intelligibility in noise would be to place information in those spectral regions least affected by the noise. However, the current study found little evidence that speakers were able to modify their speech productions in this way to take advantage of noise-free regions. In the presence of high-pass noise, speech parameters such as F0 and F1 frequencies, and spectral CoG did not shift downwards but instead increased relative to speaking in quiet conditions. One explanation for this result is that the increase in vocal effort caused by noise limited the scope for variability of other speech parameters such as fundamental frequency. However, there remains the possibility that under more realistic communicative conditions, speakers may adopt active strategies to reduce the effect of noise for listeners.

Alku, P., Vintturi, J., and Vilkman, E. (**2002**). “Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation,” Speech Commun. **38**, 321–334.

Boersma, P. (**1993**). “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a samples sound,” Proc. Inst. Phonetic Sci. **17**, 97–110.

Boersma, P., and Weenink, D. (**2005**). “Praat: Doing phonetics by computer (version 4.3.14) (computer program),” from http://www.praat.org. (Last viewed May, 2005).

Bond, Z., Moore, T., and Gable, B. (**1989**). “Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask,” J. Acoust. Soc. Am. **85**, 907–912.

Burg, J. P. (**1975**). “Maximum entropy spectrum analysis,” Ph.D. thesis, Stanford University, Palo Atto, CA.

Cooke, M. P., Barker, J., Cunningham, S., and Shao, X. (**2006**). “An audio-visual corpus for speech perception and automatic speech recognition,” J. Acoust. Soc. Am. **120**, 2421–2424.

Egan, J. P., and Hake, H. W. (**1950**). “On the masking pattern of a simple auditory stimulus,” J. Acoust. Soc. Am. **22**, 622–630.

Gramming, P., Sundberg, J., Ternstöm, S., Leanderson, R., and Perkins, W. (**1988**). “Relationship between changes in voice pitch and loudness,” J. Voice **2**, 118–126.

Garnier, M. (**2007**). “Communiquer en environnement bruyant: de l’adaptation jusqu’au forçage vocal [Communication in noisy environments: From adaptation to vocal straining],” These de Doctorat de l’Universite Paris 6.

Hansen, J. H. L. (**1996**). “Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition,” Speech Commun. **20**, 151–170.

Huber, J. E., and Chandrasekaran, B. (**2006**). “Effects of increasing sound pressure level on lip and jaw movement parameters and consistency in young adults,” J. Speech Lang. Hear. Res. **49**, 1368–1379.

Junqua, J. C. (**1993**). “The Lombard reflex and its role on human listeners and automatic speech recognizers,” J. Acoust. Soc. Am. **93**, 510–524.

Junqua, J. C., Fincke, S., and Field, K. (**1998**). “Influence of the speaking style and the noise spectral tilt on the Lombard reflex and automatic speech recognition,” in International Conference Spoken Language Proceedings, pp. 467–470.

Letowski, T., Frank, T., and Caravella, J. (**1993**). “Acoustical properties of speech produced in noise presented through supra-aural earphones,” Ear Hear. **14**, 332–338.

Lombard, E. (**1911**). “Le signe de l’elevation de la voix (The sign of the rise in the voice),” Ann. Maladiers Oreille, Larynx, Nez, Pharynx **37**, 101–119.

Lu, Y., and Cooke, M. P. (**2008**). “Speech production modifications produced by competing talkers, babble and stationary noise,” J. Acoust. Soc. Am. **124**, 3261–3275.

Mokbel, C. (**1992**). “Reconnaissance de la parole dans le bruit: Bruitage/debruitage [Voice recognition in noisy environments: Sound/denoising],” Ph.D. thesis, Ecole Nationale Superieure des Telecommunications, Paris.

Pick, H. L., Jr., Siegel, G. M., Fox, P. W., Garber, S. R., and Kearney, J. K. (**1989**). “Inhibiting the Lombard effect,” J. Acoust. Soc. Am. **85**, 894–900.

Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C., and Yuchtman, M. (**1985**). “Some acoustic-phonetic correlates of speech produced in noise,” in International Conference on Acoustics Speech and Signal Processing, pp. 1581–1584.

Schulman, R. (**1985**). “Dynamic and perceptual constraints of loud speech,” J. Acoust. Soc. Am. **178**, S37.

Stevens, K. N. (**2000**). *Acoustic Phonetics (Current Studies in Linguistics)* (MIT, Cambridge, MA).

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 analysis,” J. Acoust. Soc. Am. **84**, 917–928.

Tartter, V. C., Gomes, H., and Litwin, E. (**1993**). “Some acoustic effects of listening to noise on speech production,” J. Acoust. Soc. Am. **94**, 2437–2440.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (**1999**). *The HTK Book 2.2*, (Entropic, Cambridge).