The contribution of durational and spectral changes to the Lombard speech intelligibility benefit

Martin Cooke[a)]
Language and Speech Laboratory, Universidad del País Vasco, 01006 Vitoria, Spain

Catherine Mayo
Centre for Speech Technology Research, University of Edinburgh, Edinburgh, United Kingdom

Julián Villegas
Computer Arts Laboratory, University of Aizu, 965-8580 Aizu Wakamatsu, Japan

(Received 14 February 2013; revised 3 December 2013; accepted 17 December 2013)

Speech produced in the presence of noise (Lombard speech) is typically more intelligible than speech produced in quiet (plain speech) when presented at the same signal-to-noise ratio, but the factors responsible for the Lombard intelligibility benefit remain poorly understood. Previous studies have demonstrated a clear effect of spectral differences between the two speech styles and a lack of effect of fundamental frequency differences. The current study investigates a possible role for durational differences alongside spectral changes. Listeners identified keywords in sentences manipulated to possess either durational or spectral characteristics of plain or Lombard speech. Durational modifications were produced using linear or nonlinear time warping, while spectral changes were applied at the global utterance level or to individual time frames. Modifications were made to both plain and Lombard speech. No beneficial effects of durational increases were observed in any condition. Lombard sentences spoken at a speech rate substantially slower than their plain counterparts also failed to reveal a durational benefit. Spectral changes to plain speech resulted in large intelligibility gains, although not to the level of Lombard speech. These outcomes suggest that the durational increases seen in Lombard speech have little or no role in the Lombard intelligibility benefit. © *2014 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4861342]

PACS number(s): 43.70.Mn, 43.71.Gv, 43.72.Dv [AA] Pages: 874–883

I. INTRODUCTION

Speech produced in the presence of noise—Lombard speech (LS)—displays a large number of acoustic differences from speech produced in quiet—"plain" speech.[1] Compared to plain speech, LS typically exhibits increases in fundamental frequency ($F0$) mean and range, a flatter spectral tilt, and a slower speech rate (e.g., Summers *et al.*, 1988; Junqua, 1993; Garnier *et al.*, 2006). LS has also been found to be more intelligible than plain speech when presented in noise, even when the difference in intensity between the two styles is removed (Dreher and O'Neill, 1957; Summers *et al.*, 1988; Pittman and Wiley, 2001). Of course, LS itself is not especially prevalent in day-to-day speech communication. Rather, the study of LS is motivated by a desire to understand the basis for its intelligibility-enhancing properties. Indeed, speech modification algorithms inspired by LS for both synthetic (Langner and Black, 2005; Valentini-Botinhao *et al.*, 2012) and natural speech (Skowronski and Harris, 2006; Zorila *et al.*, 2012) have been shown in a recent evaluation to lead to gains worth more than 4 dB of additional noise (Cooke *et al.*, 2013).

However, it is not clear which of the acoustic-phonetic characteristics of LS might be related to this intelligibility gain: There have been very few studies examining the relationship between intelligibility and the intonational, spectral, and durational features of LS. One of these studies, carried out by Pittman and Wiley (2001), found correlations between LS intelligibility and LS vocal level and spectral composition. However, the authors note that these correlations were small and inconsistent. Pittman and Wiley suggest that the LS-related intelligibility gain might be the result of complex interactions between the acoustic characteristics of LS, rather than a simple one-to-one relationship between individual parameters and intelligibility.

In an effort to clarify these possible interactions, Lu and Cooke (2008, 2009) carried out a number of studies investigating the relative influence of various global LS characteristics on intelligibility. Lu and Cooke (2008) demonstrated that LS intelligibility is well-predicted by a model of energetic masking, and proposed that the higher intelligibility of LS could be due to (1) the increase in duration seen in LS, which provides more opportunities to glimpse acoustic information, and (2) the shift in energy to higher frequency regions, where it is less likely to be masked by speech-shaped noise (SSN).

Lu and Cooke (2009) investigated the effect on intelligibility of artificially manipulating spectral tilt and $F0$. Plain speech was altered to have the flattened spectral tilt and/or the raised $F0$ mean of naturally-produced LS. The intelligibility of these manipulated utterances was compared to that of un-manipulated plain and un-manipulated LS. The results demonstrated that, while artificially raising $F0$ mean does

[a)]Author to whom correspondence should be addressed. Electronic mail: m.cooke@ikerbasque.org

 0001-4966/2014/135(2)/874/10/$30.00

not significantly improve the intelligibility of plain speech, flattening spectral tilt does lead to significant intelligibility improvements, compared to unmodified plain speech. Both of these relationships hold true whether *F*0 and spectral tilt were modified separately or in combination. These results taken together suggest a primary role for spectral information in the LS intelligibility advantage.

However, there are a number of issues that Lu and Cooke (2009) did not address. First, while some of the modified speech in that study was found to be more intelligible in noise than unmodified plain speech, *none* of the modified speech was as intelligible in noise as unmodified LS. That is, there is some characteristic of LS, other than spectrum and *F*0, which contributes to its intelligibility advantage over plain speech. Since intelligibility scores correlated well with the amount of the spectro-temporal plane which escaped masking—a quantity which increases with duration—Lu and Cooke (2009) suggested a possible role for a lengthened duration in LS intelligibility. Previous studies examining the role of duration in speech intelligibility have been contradictory. Some studies of intrinsic inter-talker intelligibility differences have shown duration to be correlated with intelligibility (Bond and Moore, 1994; Hazan and Markham, 2004), while others have not (Cox *et al.*, 1987; Bradlow *et al.*, 1996). Another form of modified speech which, like LS, results in durational changes is "clear speech," i.e., speech produced in response to an instruction to speak clearly (e.g., Picheny *et al.*, 1986). Krause and Braida (2002) were able to train talkers to produce clear speech at plain speech rates (i.e., faster than naturally-produced clear speech). However, only a small, non-significant intelligibility advantage was found for fast-clear speech, suggesting that while it is not compulsory for talkers to produce clear speech with a slow speech rate, the durational aspects of clear speech may provide some intelligibility benefit for listeners. In contrast, artificially increasing the duration of plain speech (either uniformly or non-uniformly) has failed to cause significant changes in intelligibility (Schon, 1970; Schmitt, 1983; Picheny *et al.*, 1989; Uchanski *et al.*, 1996; Nejime and Moore, 1998). As yet, the relationship between intelligibility and duration in LS has not been specifically investigated.

Second, all of the acoustic modifications in Lu and Cooke (2009) were performed at a global, or per-utterance, level. However, most studies of parameter differences between LS and plain speech have demonstrated that modifications are not performed by talkers in a uniform manner across all segment types (e.g., Picheny *et al.*, 1986; Stanton *et al.*, 1988; Junqua, 1993; Lu and Cooke, 2008). Likewise, studies of clear speech have demonstrated that while vowels are generally lengthened relatively uniformly across the segment (Moon and Lindblom, 1994), albeit at different relative amounts for lax and tense vowels (Picheny *et al.*, 1986), increasing the duration of consonants often involves the reintroduction of features that are deleted or reduced in plain speech (e.g., stop consonant releases, Bradlow, 2003; Bond and Moore, 1994; Picheny *et al.*, 1986). Similarly, spectral tilt changes observed from plain speech to LS have also been found to differ across speech segments (Lu and Cooke, 2008). It may thus be the case that, in order to model the changes in duration and spectral information between plain speech and LS, local modifications would be more appropriate.

Finally, a number of previous studies point toward a possible asymmetry in the effect of clear speech characteristics. More specifically, some studies have found that removing individual clear speech characteristics from naturally-produced clear speech reduces intelligibility compared to that found for unmodified clear speech (Nejime and Moore, 1998; Schmitt, 1983; Schon, 1970; Uchanski *et al.*, 1996), while other studies have found that adding individual clear speech characteristics to plain speech does not always increase intelligibility compared to that found for plain speech (Picheny *et al.*, 1989; Uchanski *et al.*, 1996). However, to our knowledge, this asymmetry has not been investigated using matched plain and LS stimuli.

A subjective perceptual experiment was designed to address these three issues. The influence of durational and spectral features of LS on intelligibility was investigated by independently manipulating these characteristics at a global (per utterance) level and at a local (frame-based) level. Both natural plain speech and naturally-produced LS were manipulated: Plain speech was altered to have the durational and spectral characteristics of matched LS ("adding" LS characteristics), while LS was altered to have the durational and spectral characteristics of plain speech ("removing" LS characteristics). Given the absence of effect in Lu and Cooke (2009), *F*0 modifications were omitted from this experiment.

II. EXPERIMENT: INTELLIGIBILITY OF MODIFIED PLAIN AND LOMBARD UTTERANCES

A. Method

1. Participants

Twenty-six native British English listeners (12 female and 14 male) with a mean age of 23 yrs (standard deviation = 5) participated in the experiment. All listeners had normal hearing thresholds (<20 dB hearing level) in the range of 125 Hz to 8 kHz, as tested with a Kamplex KS 8 screening audiometer (London, UK). Ethical permission was obtained following the University of Edinburgh ethics procedure. Listeners were paid for their participation.

2. Plain and LS corpus

The unmodified plain and unmodified LS stimuli used in the current study were extracted from a corpus recorded by Lu and Cooke (2008) and used in the perceptual experiments conducted by Lu and Cooke (2009). In this corpus, eight native British English talkers (four male and four female) produced simple six-word sentences that conformed to the pattern used in the Grid corpus (Cooke *et al.*, 2006), such as "Set white in X 8 again" or "Lay blue in N 3 now." While Grid sentences are highly formulaic, they limit the use of higher-level linguistic knowledge which has been shown to influence the Lombard effect (Patel and Schell, 2008). The talkers recorded by Lu and Cooke (2008) produced these sentences both in quiet (plain speech) and while listening to

speech-shaped noise (SSN) at three different noise levels [82, 89, and 96 dB sound pressure level (SPL)]. Each of the 8 talkers produced a different set of 50 Grid-type sentences, for a total of 400 utterances in each talking condition. The current study made use of the plain utterances and the text-matched LS utterances recorded in SSN at 96 dB SPL and sampled at 25 kHz.

Figure 1 illustrates durational and spectral tilt values for text-matched plain and LS pairs of sentences. Spectral tilt was computed as the slope of the linear regression of the log-energies in each band of a third-octave filterbank. Diagonal lines indicate the points at which plain speech and LS values were identical: Any points which fall along this line represent pairs of utterances in which the talker did not differ from plain speech to LS in their use of spectral or durational settings. Points that fall above the diagonal line indicate pairs of utterances in which the talker produced LS with a flatter spectral tilt or longer sentence duration than the matched plain speech utterance, which is the expected behavior given previous studies of LS. Points that fall below the diagonal line indicate pairs of utterances in which the speaker produced LS with a steeper spectral tilt or shorter sentence duration than the matched plain speech utterance. On average, talkers produced LS utterances that were 7% slower and with a spectral tilt that was 0.37 dB/octave flatter than those found for the text-matched plain speech utterances. Around 90% of LS utterances had shallower spectral tilt than their plain speech counterparts, while 75% of LS utterances were longer than their plain counterparts. Note that these data are for entire sentences; for individual consonants and vowels durational and tilt changes are not uniform, as shown for the Grid sentences in Lu and Cooke (2008).

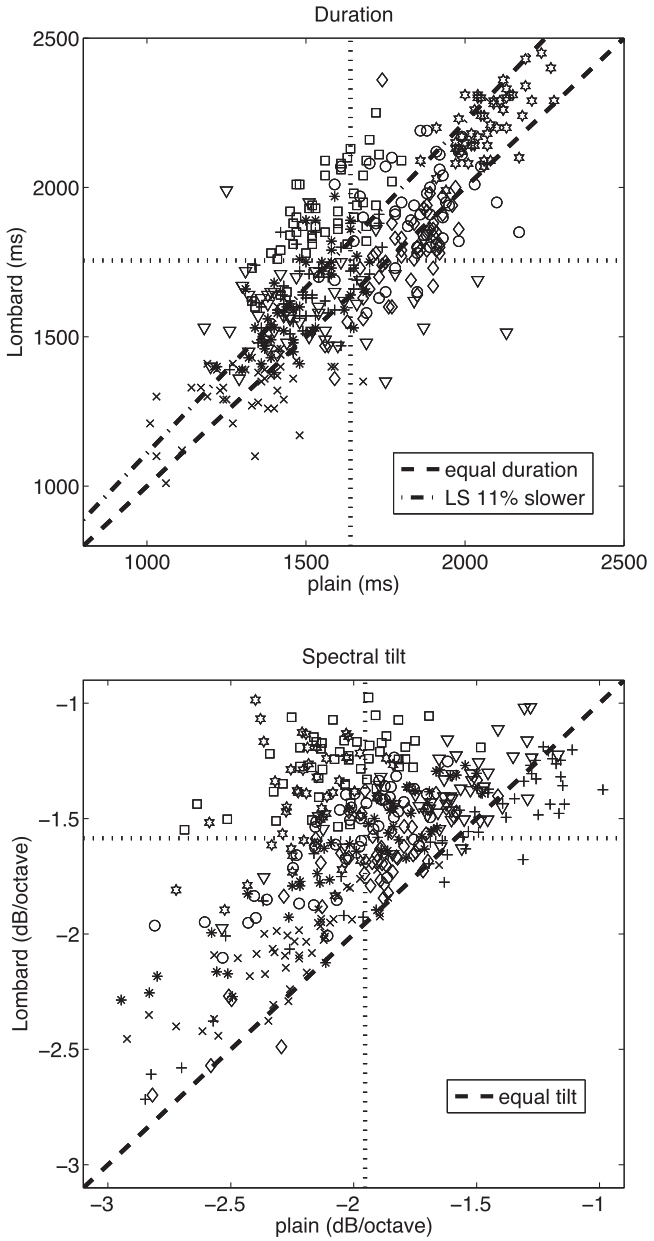

FIG. 1. Scatterplots of duration and spectral tilt for sentences in plain and LS. Each sentence pair is represented by a single symbol, with utterances produced by the same speaker grouped by symbol. The dotted horizontal and vertical lines are at the means of each sample while the diagonal line indicates where both plain and LS sentences had equal values of the characteristic. In the duration plot, values above the line marked "LS 11% slower" are members of the LS-SLOWER set.

3. Stimuli

In addition to the plain and Lombard speech eight sets of modified speech were also created for the current study made up of all combinations of the following manipulations: (1) Two acoustic attributes—duration and spectrum—were changed independently; (2) for each manipulation, changes were made in two directions: Either plain speech was modified to have the relevant characteristics of LS, or LS was modified to have the relevant characteristics of plain speech; and (3) modifications were made at two scales, global or local, as explained below. Modifications were carried out based on parameters extracted from pairs of same-text utterances, i.e., plain speech and its Lombard counterpart. Modifications were performed at the full spectral bandwidth.

Global durational modification was via linear expansion or contraction to match each utterance with its counterpart, conducted using pitch-shift overlap-add (PSOLA) as implemented in Praat (Boersma and Weenink, 2013). The periodicity detection used in duration modifications requires a defined minimum and maximum $F0$ in order to operate adequately. These values were manually adjusted independently for each speaker across all sentences so no audible artifacts (e.g., octave jumping) were detected in any modified utterance.

Global spectral modifications were achieved with the double filtering procedure used in Lu and Cooke (2009). First, the spectrum of each utterance was flattened by filtering using an 18-pole linear prediction approximation to the inverse of its spectrum. A low order approximation was used to avoid transplanting fine spectral detail (e.g., related to the individual harmonics) from one utterance to its counterpart. Second, the resulting signal was filtered with the linear predictive fit of the counterpart utterance. The overall effect is to transplant spectral information from the counterpart utterance (i.e., from LS for modifications to plain speech, and from plain speech for modifications to LS). Figure 2 depicts the outcome of this procedure for the plain and LS corpora (note that while the figure shows long-term average spectra

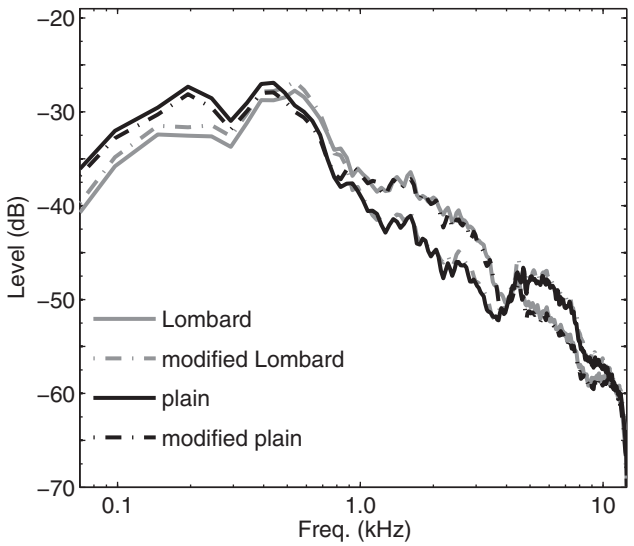

FIG. 2. Long-term average speech spectra of the plain and LS corpora (solid lines), along with similar spectra for speech with spectral modifications designed to match plain or LS.

measured across the corpus, the procedure was carried out using spectra estimated for each plain-LS utterance pair individually). In spite of the use of a low-order linear prediction spectrum, the original and transformed long-term spectra are well-matched apart from a slight mismatch of around 1 dB in the region below 400 Hz.

Local durational modifications were achieved using a combination of dynamic time warping and PSOLA techniques as implemented in the VocALign program (Synchroarts, 2011). Local durational modification is illustrated in Fig. 3, where the result of time aligning a plain utterance [panel (a)] to its counterpart LS utterance [panel (c)] is shown in panel (b).

The procedure for local spectral modifications is also illustrated in Fig. 3 for the case of modifications in the direction LS to plain (removing LS characteristics from LS). First, the plain utterance is time-aligned to its LS counterpart. Then, spectral information from the time-aligned plain utterance is mapped onto the LS utterance using the same spectral transformation approach as in the global modifications, except that rather than the transformation being performed once per utterance it is done separately in each time window. Here, 20 ms Hann-windowed frames with 50% overlap were used. The bottom panel of Fig. 3 shows the result of the local spectral modification process. Modifications from plain to LS are performed in a similar manner: The LS utterance is time-aligned to its plain counterpart prior to the transformation of LS spectral information to plain speech.

4. Procedure

The listening experiment was conducted in individual sound-attenuating booths at the University of Edinburgh. Utterances were presented to participants co-gated with a SSN masker mixed at the same signal-to-noise ratio (−9 dB) used in Lu and Cooke (2009). The SSN sample had the same long-term amplitude spectrum as that of the plain Grid sentences. Utterance-plus-noise pairs were delivered diotically under computer control through Beyerdynamic DT770 headphones.

As explained above, each of the 8 talkers produced 50 sentences, leading to a possible 400 sentences in each of the 10 conditions (plain, LS, and the 8 modifications). In each condition, the set of 400 sentences was divided into 10 blocks of 40 (8 from each of the 5 talkers) and listeners were assigned to 1 block in each condition following a Latin square design.

For the purposes of perceptual testing, the two key words in these utterances were (1) the fourth item in every sentence, which was a monosyllabic letter name ("A" through "Z" excluding "W"), and (2) the fifth item, which was a digit (0–9). Note that the Grid task also permits responses to the color keyword, but its use is largely reserved for informational masking studies where it is necessary to denote which is the target sentence in the presence of competing speech material (e.g., to report the alpha-digit keywords in the sentence containing the word "white"). Since the color keyword has only four alternatives and would lead to more complex keyboarding requirements, listeners were asked to respond only to the alpha-digit combination in the current study.

Block presentation order was counter-balanced across participants. Each stimulus was presented once, after which participants responded by selecting first the letter and then the digit keyword from an onscreen keyboard. Participants were instructed to guess when in doubt: There was no null response option. After inputting the second of the two keywords, the next stimulus was presented after a short pause. In this way, the experiment was self-paced. Participants were allowed to pause between blocks. On average participants were able to finish the whole procedure in about 30 min. Since all subjects had participated in a pilot experiment with a similar setup 2 weeks before this experiment, no practice sessions were conducted.

B. Results

Listeners correctly identified 54.4% and 70.2% of keywords in unmodified plain and unmodified LS, respectively, a Lombard gain of nearly 16 percentage points (p.p.). These values are very close to the 56% and 74% and corresponding 18 p.p. gain reported in Lu and Cooke (2009).

To permit direct comparison between the effect of adding LS characteristics to plain speech, and the effect of removing LS characteristics from LS, we calculated *changes* in correct keyword identification from the unmodified plain and LS baseline scores reported above to those seen in response to the modified speech conditions. For those conditions derived from plain speech the quantity plotted is $\text{score}_{\text{condition}} - \text{score}_{\text{plain}}$ while for modifications to LS the quantity plotted is $\text{score}_{\text{LS}} - \text{score}_{\text{condition}}$.

Using the above change-from-baseline keyword scores, separate repeated-measures analyses of variance (ANOVAs) with modification as a within-subjects factor and sentence subset (i.e., the selection of sentences assigned to each listener) as a between-subjects factor were carried out on (1) plain speech modified to have LS characteristics, and (2) LS speech modified to have plain speech characteristics. In neither case was sentence subset a significant factor [$p = 0.62$ and 0.30, respectively].

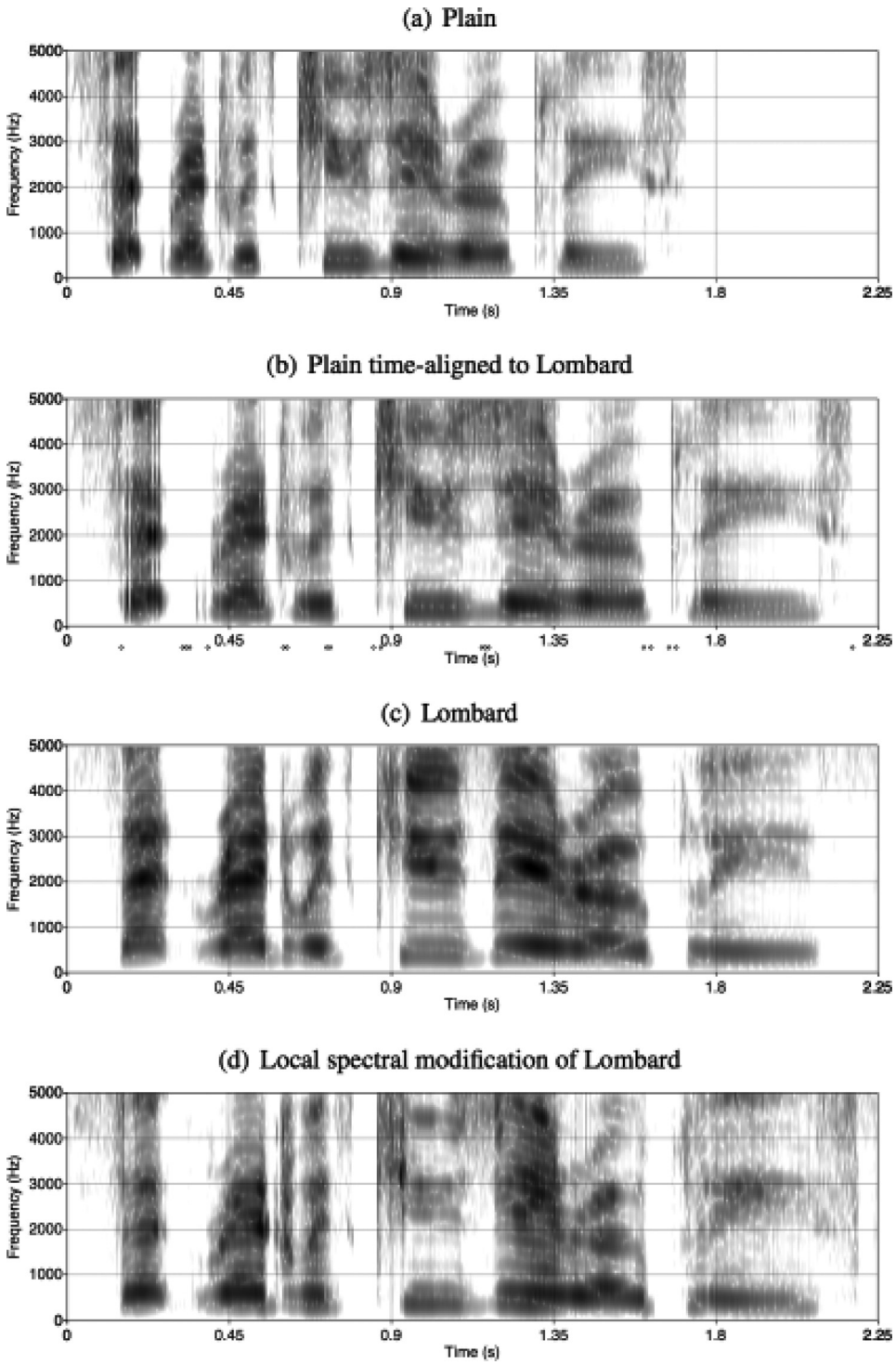

FIG. 3. Illustration of local spectral modification from LS to plain for the utterance "*Set red with Q zero please*." First, plain speech (a) is time-aligned to LS (c), producing the aligned version shown in panel (b). Second, local spectral information from the time-aligned plain speech is transplanted into LS, resulting in the modified speech shown in panel (d). Dots under the plain time-aligned to Lombard plot show the location of frames where the difference between the modification and the target style is larger than 9 dB.

Figure 4 plots change-in-intelligibility scores relative to plain and LS baselines. Both directions of modification show a clear main effect of type of modification: Plain with LS characteristics added [$F(4, 64) = 48.7$, $p < 0.001$, $\eta^2 = 0.59$; Fisher's LSD: 3.3 p.p.], LS with LS characteristics removed [$F(4, 64) = 43.5$, $p < 0.001$, $\eta^2 = 0.53$; Fisher's LSD: 3.2 p.p.].

The global application of LS spectral characteristics to plain speech increased the intelligibility of that speech almost to the level of that seen for unmodified LS, falling short by around 3 p.p., while globally adding LS durational changes had no significant impact on intelligibility, with scores around 2 p.p. lower than those in the plain speech condition. Looking at global modifications in the other direction, globally removing LS spectral characteristics from LS significantly reduced the intelligibility of that speech by just over 10 p.p. but not to the level of unmodified plain speech, while globally removing LS durational characteristics from LS led to a small but significant loss in intelligibility of 5.6 p.p. relative to unmodified LS.

Local modifications produced a somewhat different set of results. While local spectral changes to plain speech were beneficial, the size of the increase in scores fell around 3.5 p.p. short

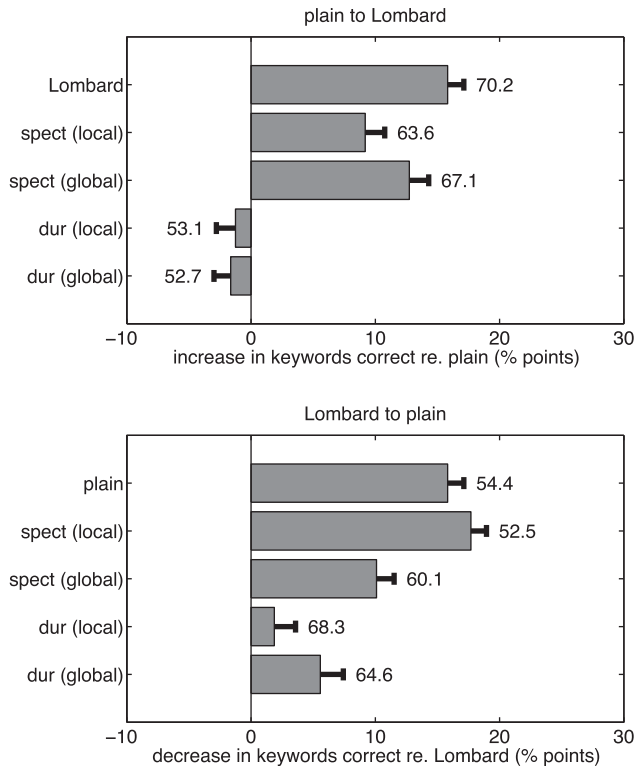

FIG. 4. Changes in keywords correct scores. The upper panel shows p.p. gains in scores relative to plain speech for modifications made to plain speech, along with unmodified LS, while the lower panel shows losses relative to LS for modifications made to LS, along with unmodified plain speech. Error bars here and elsewhere correspond to 1 standard error. The absolute keyword scores are also given.

of that seen for global changes (that is, local spectral changes did not increase speech intelligibility as much as did global spectral changes), a difference larger than Fisher's LSD. Local durational changes to plain speech had no effect. Local spectral changes to LS resulted in a very large decrease in scores to a level equivalent to that of plain speech, while local durational changes in this direction had no detrimental effect.

Further ANOVAs for the modified speech conditions only (i.e., excluding the plain and LS baselines) with factors of parameter type (spect, dur) and scale of modification (local, global) confirmed these findings. For modifications to plain speech, there was an effect of parameter type [$F(1, 25) = 137$, $p < 0.001$, $\eta^2 = 0.40$], but no significant effect of modification scale [$p = 0.27$] nor an interaction between parameter and scale [$p = 0.09$]. However, the parameter-scale interaction was statistically significant for modifications made to LS [$F(1, 25) = 19.2$, $p < 0.001$, $\eta^2 = 0.11$], as seen in the lower panel of Fig. 4.

C. Discussion

The finding reported in Lu and Cooke (2009) that global spectral characteristics of LS play a relatively large role in the LS-related intelligibility gain over plain speech is replicated here: Globally adding LS spectral characteristics to plain speech induced a near 13 p.p. increase in intelligibility. The current study extends this finding to the effect of making spectral changes in the reverse direction. Applying spectral characteristics of plain speech to LS decreased intelligibility by 10 p.p. The results of the experiment additionally shed light on the intelligibility-enhancing role of LS spectral characteristics at the level of local spectral changes. Local changes to plain speech were also beneficial, but by slightly less than global changes, while local changes to LS produced a far larger decrease in scores, more than wiping out the original Lombard benefit. We discuss possible reasons for this asymmetry in Sec. IV below.

It initially appears that the durational characteristics of LS have a very small or negligible effect on intelligibility, with three of the four durational modifications resulting in no significant change in listeners' ability to correctly identify keywords. Adding LS durational characteristics to plain speech did not improve intelligibility, whether done at a global or a local level, and locally removing LS characteristics did not decrease intelligibility. However, removing LS durational characteristics at a global level significantly worsened intelligibility compared to that seen for unmodified LS. Interestingly, although not statistically significant, the small change that was observed when applying LS durational characteristics to plain speech was in the direction of *worsened* intelligibility, rather than in the expected direction.

The absence of a significant effect of durational changes in three of the four conditions raises questions about the durational characteristics of the stimuli used in the current study. As noted above, the expected direction of durational change made by talkers when shifting from plain speech to LS is to elongate: Simply put, talkers tend to produce LS more slowly than they produce plain speech (Junqua, 1993; Lu and Cooke, 2008). However, it is also the case that individual talkers can differ greatly in the acoustic-phonetic strategies they adopt to produce LS (Summers *et al.*, 1988; Junqua, 1996). Indeed, as shown earlier in Fig. 1, around 25% of the plain speech utterances used in the current study were produced more slowly than their LS counterparts, raising the possibility that the absence of an overall durational benefit of LS is due to the presence of these "Lombard-faster" utterances. To address this issue, Sec. III presents the results of re-analyzing the perceptual effects for a subset of utterances where the LS member of the pair is clearly slower than its plain speech counterpart.

III. INTELLIGIBILITY BENEFITS FOR UTTERANCES WITH LONGER LS DURATIONS

A. Subset selection

Since many plain and LS utterance pairs have similar durations, as is evident from the clustering of utterances near the diagonal in Fig. 1, the selection of utterances pairs based on a strict "Lombard-slower" criterion results in the inclusion of many pairs with relatively small durational differences. Therefore, in order to promote the emergence of a putative LS durational benefit, only those utterances with substantially longer LS durations were selected. Specifically, the analysis was based on those utterance pairs in the upper tercile of durational change, corresponding to LS utterances whose duration was 111% to 159% of that of their plain counterparts. We refer to this subset as the LS-SLOWER group.

These sentences are represented by the points above the upper diagonal in the duration panel of Fig. 1.

B. Results

Keyword scores for the LS-SLOWER group were 52.6% and 75.5% for plain and LS, respectively, a Lombard advantage of nearly 23 p.p. This gain is larger than the 16 p.p. overall Lombard benefit reported in Sec. II. Differences in keyword scores from plain and LS baselines for the LS-SLOWER group are plotted in Fig. 5. In general, the LS-SLOWER scores are always larger (by 3 to 7 p.p.) than those for the complete set of utterances as seen earlier in Fig. 4, but follow a very similar pattern as a function of type of modification. Interestingly, changes over baselines were also higher for *spectral* modifications—in which duration did not change—than was the case for the complete corpus.

The effect of sentence subset was non-significant for the LS-SLOWER group [$p = 0.39$ and 0.55 for plain-to-LS and LS-to-plain, respectively]. Manipulation was statistically-significant [plain-to-LS: $F(4, 64) = 29.0$, $p < 0.001$, $\eta^2 = 0.44$; LS-to-plain: $F(4, 64) = 17.8$, $p < 0.001$, $\eta^2 = 0.37$] with Fisher's LSD values of 4.6 and 5.5 p.p., respectively. Local and global durational modifications were statistically equivalent. The modest increases in keyword scores seen for duration in the plain to LS direction were not statistically significant. For durational modifications made to LS, global modifications led to a significant decrease in keyword scores, with a similar but marginally non-significant tendency for local modifications.

Further ANOVAs for the modified speech conditions only with factors of parameter type and scale of modification bore out these findings. For modifications to plain speech, there was an effect of parameter type [$F(1, 25) = 104$, $p < 0.001$, $\eta^2 = 0.26$], but no significant effect of modification scale [$p = 0.91$] nor an interaction between parameter and scale [$p = 0.21$]. The parameter-scale interaction was statistically significant for modifications made to LS [$F(1, 25) = 7.3$, $p < 0.05$, $\eta^2 = 0.05$].

Perhaps the most striking outcome of the re-analysis is the difference in gains for the non-durational (i.e., spectral) manipulations. It seems reasonable to assume on the basis of the significantly larger changes in scores in response to spectral manipulation to LS-SLOWER speech that these utterances possessed spectral characteristics which conferred greater benefits in the face of masking noise. Indeed, the size of the overall Lombard benefit for the LS-SLOWER subset was 7 p.p. higher than that observed for the entire corpus. This notion is supported by data shown in Fig. 6, which plots changes between LS and plain speech in spectral tilt against changes in duration across the entire corpus. The moderate positive correlation [$\rho = 0.34$, $p < 0.001$] indicates a tendency for utterances whose speech rate is slower in LS to have a flatter average spectrum. For the LS-SLOWER subset the correlation is slightly higher [$\rho = 0.36$, $p < 0.001$].

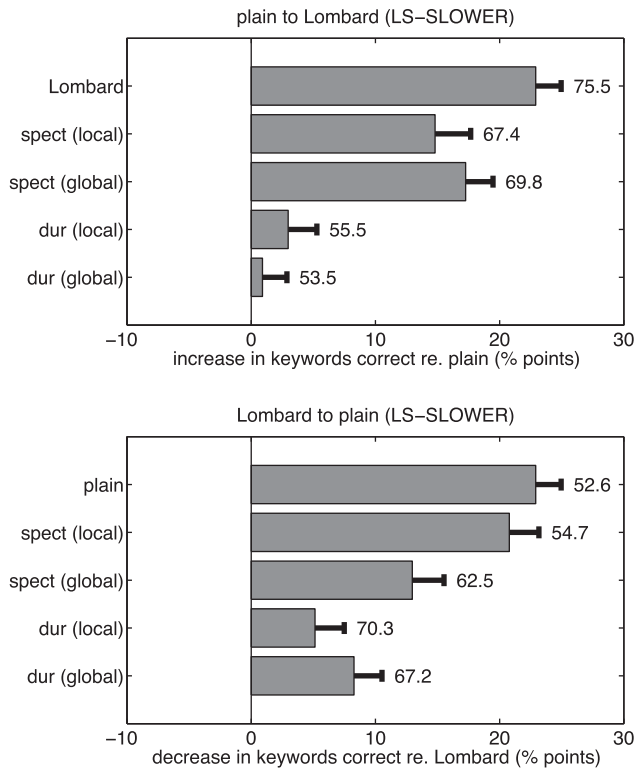

FIG. 5. Effects of modifications for utterances pairs where the LS member is slower than its plain counterpart.

C. Glimpsing analysis

While spectral tilt provides a useful scalar approximation to the spectrum, it allows only a crude estimate of the effect of energetic masking by SSN. A more direct estimate of the amount of spectro-temporal information escaping the masker can be obtained by a glimpsing analysis (Cooke, 2006), which measures the percentage of time-frequency regions in an auditory representation where speech contains more energy than the masker. Here, glimpses were derived from modeled spectro-temporal excitation patterns computed

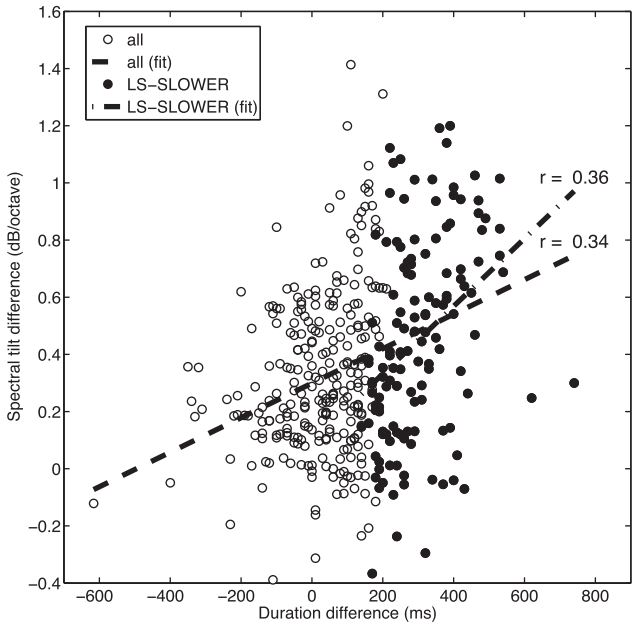

FIG. 6. Durational and tilt differences between Lombard and plain speech. Each point represents a single utterance pair. Note that the set marked "all" includes the LS-SLOWER subset (i.e., both open and filled circles). Likewise, the best fit for "all" also includes the LS-SLOWER subset.

by separate analysis of speech and masker signals using a 58 channel gammatone filterbank with center frequencies spanning the range 100 to 8000 Hz, followed by extraction of the Hilbert envelope, temporal integration, downsampling to 10 ms frames and log compression.

The percentage of time-frequency regions where the speech excitation pattern exceeds that of the masker, averaged across the corpus, is presented for each condition in Fig. 7. Glimpse percentage is highly-correlated with keyword scores both for the overall corpus [$\rho = 0.93$, $p < 0.001$] and for the LS-SLOWER subset [$\rho = 0.92$, $p < 0.001$]. LS has 48% more glimpses than plain speech across the entire corpus, a value that increases to 58% for the LS-SLOWER subset. Since glimpse *percentage* is independent of durational changes, these figures demonstrate that although the LS-SLOWER subset was based on durational changes with respect to plain speech, energetic masking differences play the dominant role in the increased LS advantage for this group.

D. Interim discussion

No significant benefit of manipulations which resulted in slower speech (i.e., in the plain to LS direction) was found for global nor local modifications. On the other hand, manipulations resulting in faster speech (i.e., in the LS to plain direction) led to significant decreases in keyword scores. These findings echo and bring into sharper focus the results of the analysis of the entire corpus presented in Sec. II, and suggest that any contribution made by a slower speech rate to the Lombard benefit in noise is very limited.

IV. GENERAL DISCUSSION

We conclude by examining how the findings of the current study inform the three main research questions outlined in Sec. I.

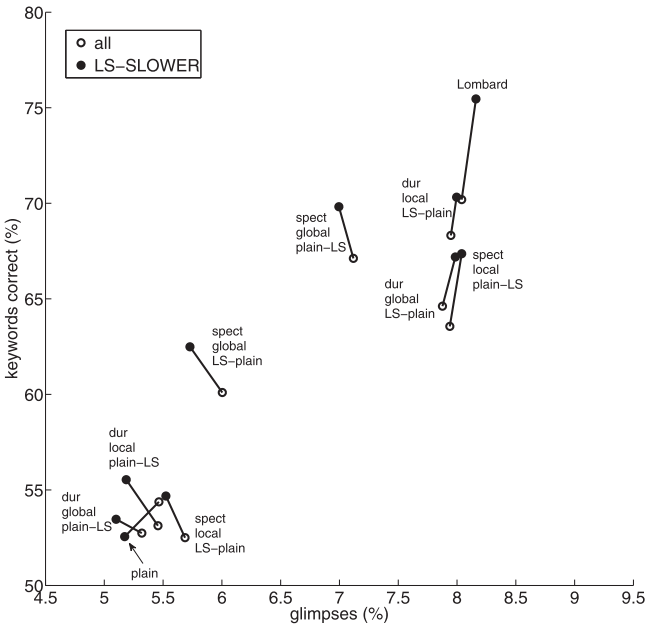

FIG. 7. Intelligibility as a function of glimpse proportion. Open circles are based on scores for the entire corpus while filled circles denote scores based solely on the LS-SLOWER subset.

A. The role of duration in the LS intelligibility benefit

The principal finding of Lu and Cooke (2009), that applying spectral characteristics of LS to plain speech results in a large intelligibility gain (though not quite to the level of LS), was confirmed in the current study. However, their speculation that durational differences between plain and LS might act to take intelligibility up to the level of LS was not supported. Note that the speech material upon which Lu and Cooke (2009) based their hypothesis corresponds to the undifferentiated set of stimuli presented in Sec. II, where the modest change in keyword intelligibility scores as a result of durational manipulations to plain speech was in the opposite direction to the anticipated increase in intelligibility. Even for responses to stimuli where the LS utterance was clearly slower than its plain counterpart utterance, increases in intelligibility were small and not statistically-significant. In this respect, our findings agree with those of non-LS studies which failed to find a significant benefit of increasing the duration of plain speech (Schon, 1970; Schmitt, 1983; Picheny *et al.*, 1989; Uchanski *et al.*, 1996; Nejime and Moore, 1998).

The lack of benefit of a slower speech rate in LS could be due to a number of factors. One is that the amount of slowing down observed in LS may be too modest to produce gains in performance. In the current study, the upper tercile of LS durational increase corresponded to changes in the range 11% to 59%, with a mean increase of 20%. For comparison, when talkers are asked to speak clearly they typically produce utterances at around 100 words/min, far lower than the 180 words/min of conversational speech (Krause and Braida, 2002).

It is worth noting that intelligibility scores measure the *net* effect of stimulus manipulation, which could conceivably result from the combination of both beneficial and harmful factors. For example, it is possible that the lack of intelligibility gains associated with speech rate changes could be due to antagonistic interactions with phonological contrasts involving durational cues. An example is the use of durational differences in the preceding vowel in influencing whether a following plosive consonant is perceived as voiced or voiceless in English. Indeed, Sankowska *et al.* (2011) measured a reduced durational contrast to plosive voicing in LS compared to plain or foreigner-directed speech, highlighting the possibility that manipulations involving duration can lead to stimuli with impoverished speech cues. The notion that LS is not necessarily intrinsically beneficial (i.e., apart from its resilience to energetic masking) is further strengthened by the finding that, when presented in quiet conditions to non-native listeners, LS results in more errors than plain speech (Cooke and García Lecumberri, 2012).

B. Global versus local modifications

It might be expected that *linear* mapping of durational differences between plain and LS would exacerbate the negative interactions between Lombard durational changes and durational cues to phonemic distinctions of the kind reported by Sankowska *et al.* (2011). However, in the current study non-linear durational mapping via time alignment produced

changes in keyword scores broadly equivalent to those resulting from linear mapping. As mentioned above, it is possible that the speech rate changes were not large enough to produce distinctive local and global effects for durational manipulations.

By contrast, local and global spectral manipulations had different effects on intelligibility. Global changes to plain speech were equivalent or slightly more beneficial than local changes, while local changes to LS resulted in significantly lower intelligibility than in the global case. This outcome is contrary to our hypothesis that local changes are needed to adequately model the differences between plain and LS. The harmful impact of local spectral modifications may be the result of attempting to solve a correspondence problem between two different speech styles. If utterances in one or another style have highly reduced or missing spectral cues, this problem is ill-posed, not least because the time-warping itself (a necessary first step for local spectral modification) is likely to be less accurate. An example can be seen in Fig. 3, where the evidence for a plosive release in "please" is far stronger in the plain utterance than the Lombard utterance. An analysis of frame energies in time-aligned utterance pairs suggests the scale of the problem: Around 6% of frames possess energy differences of 9 dB or more (the location of these frames is indicated by dots in the second panel of Fig. 3).

C. Asymmetry of intelligibility changes

In general, the intelligibility gains for modifications applied to plain speech were smaller than losses incurred for modifications in the reverse direction. This extends to LS, a similar finding of asymmetry of effect evident from analysis of modifications to plain and clear speech (Nejime and Moore, 1998; Schmitt, 1983; Schon, 1970; Uchanski *et al.*, 1996). The asymmetry is particularly marked for duration, but is also evident for *local* spectral changes, though not for global spectral modifications. In the latter case, global spectral modification is equivalent to processing speech through a time-invariant filter, so it is perhaps not surprising that the net benefit in terms of reduced energetic masking is cancelled out when the filter is applied in the opposite direction.

The asymmetric effect of certain modifications can be explained if we assume that there is an intelligibility loss associated with carrying out the modification, perhaps due to antagonistic interaction with acoustic cues to phonemic distinctions of the kind suggested in Sankowska *et al.* (2011), and that the observed change in intelligibility then results from the combination of this intelligibility loss with an intrinsic modification-specific effect (which may be either a gain or loss in intelligibility depending on the direction of the modification). For example, it might be that local spectral changes to plain speech are intrinsically-beneficial, leading to a gain of α p.p., but incur a loss of β p.p. as a result of transformational factors such as those referred to above. The observed gain is then $\alpha - \beta$. However, in the opposing direction the loss is $-\alpha - \beta$, a difference of $2 * \beta$, giving rise to the observed asymmetric effect.

We have already referred to one possible cause of an intelligibility loss for local spectral changes and the lack of transformation-related loss in the case of global spectral changes, whose form is particularly simple. Concerning duration, for the LS-SLOWER subset it is notable that the largest asymmetry is evident for global changes. Given the potential effect on phonemic distinctions resulting from inappropriate changes to, for example, voice-onset time or semivowel transition duration, it is plausible that global durational modifications incur a larger transformation-related intelligibility loss than local changes.

V. CONCLUSIONS

Mapping the durational properties of LS on to plain speech produced no significant increase in keyword identification scores in sentences presented in stationary noise, regardless of whether duration was modified by linear stretching or compression, or nonlinearly via time-alignment. Mapping spectral information from Lombard to plain speech produced an increase in intelligibility which fell short of that of LS itself. These findings suggest that the LS intelligibility benefit is largely, but not wholly, due to spectral differences between plain and LS, and that durational differences have little or no role.

ACKNOWLEDGMENTS

This work was partially funded by the Listening Talker (LISTA) project, supported by the Future and Emerging Technologies (FET) program within the Seventh Framework Programme for Research of the European Commission, under FET-Open Grant No. 256230.

[1]We follow Bradlow and Alexander (2007) in adopting the term "plain speech."

Boersma, P., and Weenink, D. (**2013**). "Praat: Doing phonetics by computer," www.praat.org (Last viewed February 2013).

Bond, Z. S., and Moore, T. J. (**1994**). "A note on the acoustic-phonetic characteristics of inadvertently clear speech," Speech Commun. **14**, 325–337.

Bradlow, A. R. (**2003**). "Confluent talker- and listener-related forces in clear speech production," in *Laboratory Phonology*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin and New York), Vol. 7, pp. 241–273.

Bradlow, A. R., and Alexander, J. A. (**2007**). "Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners," J. Acoust. Soc. Am. **121**, 2339–2349.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech. I: Global and fine-grained acoustic-phonetic characteristics," Speech Commun. **20**, 255–272.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

Cooke, M., 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.

Cooke, M., and García Lecumberri, M. L. (**2012**). "The intelligibility of Lombard speech for non-native listeners," J. Acoust. Soc. Am. **132**, 1120–1129.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., and Tang, Y. (**2013**). "Evaluating the intelligibility benefit of speech modifications in known noise conditions," Speech Commun. **55**, 572–585.

Cox, R. M., Alexander, G. C., and Gilmore, C. (**1987**). "Intelligibility of average talkers in typical listening environments," J. Acoust. Soc. Am. **81**, 1598–1608.

Dreher, J., and O'Neill, J. (**1957**). "Effects of ambient noise on speaker intelligibility for words and phrases," J. Acoust. Soc. Am. **29**, 1320–1323.

Garnier, M., Bailly, L., Dohen, M., Welby, P., and Loevenbruck, H. (**2006**). "An acoustic and articulatory study of Lombard speech: Global effects on the utterance," in *Proceedings of Interspeech*, Pittsburgh, PA, pp. 2246–2249.
Hazan, V., and Markham, D. (**2004**). "Acoustic-phonetic correlates of talker intelligibility for adults and children," J. Acoust. Soc. Am. **116**, 3108–3118.
Junqua, J. (**1993**). "The Lombard reflex and its role on human listeners and automatic speech recognizers," J. Acoust. Soc. Am. **93**, 510–524.
Junqua, J.-C. (**1996**). "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," Speech Commun. **20**, 13–22.
Krause, J. C., and Braida, L. D. (**2002**). "Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility," J. Acoust. Soc. Am. **112**, 2165–2172.
Langner, B., and Black, A. W. (**2005**). "Improving the understandability of speech synthesis by modeling speech in noise," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, pp. 265–268.
Lu, Y., and Cooke, M. (**2008**). "Speech production modifications produced by competing talkers, babble, and stationary noise," J. Acoust. Soc. Am. **124**, 3261–3275.
Lu, Y., and Cooke, M. (**2009**). "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," Speech Commun. **51**, 1253–1262.
Moon, S., and Lindblom, B. (**1994**). "Interaction between duration, context, and speaking style in English stressed vowels," J. Acoust. Soc. Am. **96**, 40–55.
Nejime, Y., and Moore, B. C. J. (**1998**). "Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss," J. Acoust. Soc. Am. **103**, 572–576.
Patel, R., and Schell, K. W. (**2008**). "The influence of linguistic content on the Lombard effect," J. Speech Lang. Hear. Res. **51**, 209–220.
Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," J. Speech Hear. Res. **29**, 434–446.
Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1989**). "Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech," J. Speech Hear. Res. **32**, 600–603.
Pittman, A. L., and Wiley, T. L. (**2001**). "Recognition of speech produced in noise," J. Speech Lang. Hear. Res. **44**, 487–496.
Sankowska, J., García Lecumberri, M. L., and Cooke, M. (**2011**). "Interaction of intrinsic vowel and consonant durational correlates with foreigner directed speech," Poznan Studies Contemp. Ling. **47**, 109–119.
Schmitt, J. F. (**1983**). "The effects of time compression and time expansion on passage comprehension by elderly listeners," J. Speech Hear. Res. **26**, 373–377.
Schon, T. D. (**1970**). "The effects on speech intelligibility of time-compression and expansion on normal-hearing, hard of hearing, and aged males," J. Aud. Res. **10**, 263–268.
Skowronski, M. D., and Harris, J. G. (**2006**). "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," Speech Commun. **48**, 549–558.
Stanton, B., Jamieson, L., and Allen, G. (**1988**). "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, pp. 331–334.
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.
Synchroarts (**2011**). "Vocalign project," www.synchroarts.com (Last viewed February 2013).
Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., and Durlach, N. I. (**1996**). "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate," J. Speech Hear. Res. **39**, 494–509.
Valentini-Botinhao, C., Yamagishi, J., and King, S. (**2012**). "Speech intelligibility enhancement for HMM-based synthetic speech in noise," in *Workshop on Statistical and Perceptual Audition*, Portland, OR.
Zorila, T., Kandia, V., and Stylianou, Y. (**2012**). "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proceedings of Interspeech*, Portland, OR.