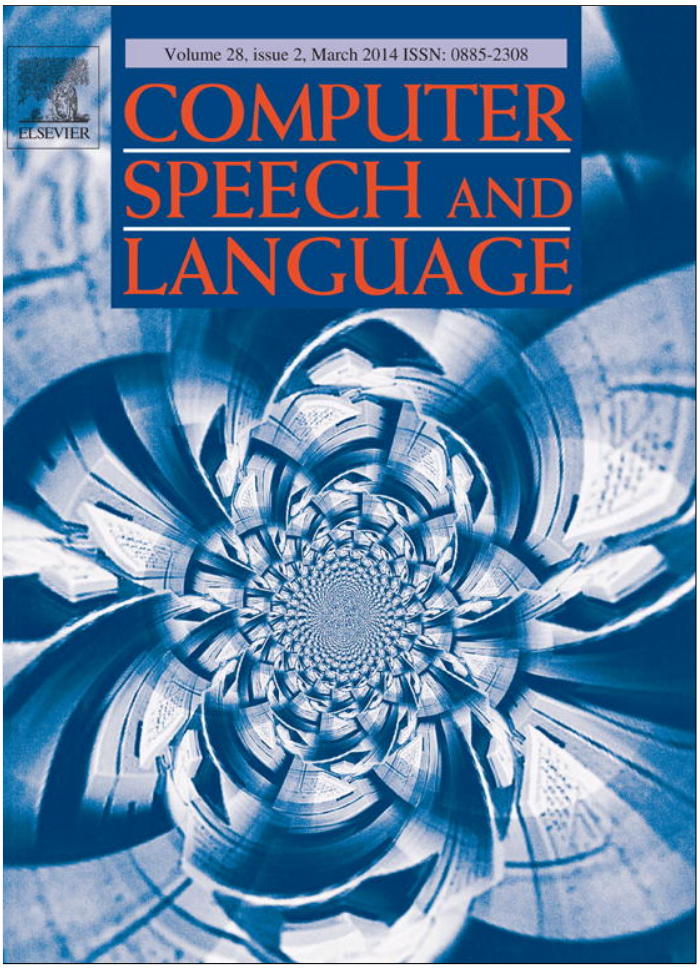

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 629–647

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Approaching speech intelligibility enhancement with inspiration from Lombard and Clear speaking styles☆

Elizabeth Godoy [a,*], Maria Koutsogiannaki [a,b], Yannis Stylianou [a,b]

[a] *Institute of Computer Science, Foundation for Research and Technology Hellas, Crete, Greece*
[b] *Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece*

Received 15 October 2012; received in revised form 5 August 2013; accepted 11 September 2013
Available online 4 October 2013

Abstract

Lombard and Clear speech represent two acoustically and perceptually distinct speaking styles that humans employ to increase intelligibility. For Lombard speech, increased spectral energy in a band spanning the range of formants is consistent, effectively augmenting loudness, while vowel space expansion is exhibited in Clear speech, indicating greater articulation. On the other hand, analyses in the first part of this work illustrate that Clear speech does not exhibit significant spectral energy boosting, nor does the Lombard effect invoke an expansion of vowel space. Accordingly, though these two acoustic phenomena are largely attributed with the respective intelligibility gains of the styles, present analyses would suggest that they are mutually exclusive in human speech production. However, these phenomena can be used to inspire signal processing algorithms that seek to exploit and ultimately compound their respective intelligibility gains, as is explored in the second part of this work. While Lombard-inspired spectral shaping has been shown to successfully increase intelligibility, Clear speech-inspired modifications to expand vowel space are rarely explored. With this in mind, the latter part of this work focuses mainly on a novel frequency warping technique that is shown to achieve vowel space expansion. The frequency warping is then incorporated into an established Lombard-inspired Spectral Shaping method that pairs with dynamic range compression to maximize speech audibility (SSDRC). Finally, objective and subjective evaluations are presented in order to assess and compare the intelligibility gains of the different styles and their inspired modifications.

Keywords: Lombard effect; Clear speech; Intelligibility enhancement

1. Introduction

In real-world communications, speakers and listeners are often immersed in noisy environments that influence both speech production and intelligibility. When faced with adverse communication conditions, human beings physically alter their manner of speaking in order to make their speech more intelligible. For example, humans adopt "Lombard" or "Clear" speaking styles, depending respectively on whether or not they are immersed in a noisy environment. Just as humans adopt their speech production to increase intelligibility, speech enhancement techniques seek to modify the speech signal in order to make it more intelligible for human listeners. With growing numbers of applications using

☆ This paper has been recommended for acceptance by S. King.

* Corresponding author. Tel.: +30 6943851605.
E-mail address: godoyec@gmail.com (E. Godoy).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.09.007

speech technologies in commercial (e.g., mobile telephone, GPS, customer service systems), military (e.g., Air Force, Ground troop relays) and medical (e.g., assisted-speech) contexts, modifications that help "speaking-devices" to be more intelligible (and consequently, relevant) are currently in high demand. Considering the intelligibility gains of the human speaking styles, acoustic phenomena observed in Lombard and Clear speech can be used to inspire speech signal modifications for intelligibility enhancement. Indeed, this is the approach adopted in the present work.

Though Lombard and Clear speech are both highly intelligible, the styles are perceptually and acoustically distinct. To begin with, the "Lombard" effect refers to the ways in which humans reflexively modify their speech when speaking in a noisy environment (Lombard, 1911). Perceptually, Lombard speech can be described as "tense" and "loud," with increased vocal effort. Compared its "normal" counterpart, Lombard speech incorporates many prosodic and segmental acoustic-phonetic modifications (Summers et al., 1988; Hanley and Steer, 1949; Dreher and O'Neill, 1957; Junqua, 1993; Womack and Hansen, 1996; Garnier et al., 2006; Lu and Cooke, 2008, 2009; Lu, 2009; Davis and Kim, 2012). Observations from these Lombard-related works include: decreased speaking rate, increased pitch, higher energy, decreased spectral tilt or spectral "flattening," as well as formant shifts particularly for F1 and F2, formant bandwidth reduction and vowel-to-consonant energy re-distribution. Among these observations, the Lombard increase in intelligibility has been shown to be largely attributed to spectral modifications (Lu and Cooke, 2009), particularly increased spectral energy in an inclusive formant band or, otherwise stated, a decreased spectral "tilt."

Alternatively to the Lombard effect, "Clear" speech strategies are adopted when a speaker in a quiet environment addresses a listener facing a communication barrier. Perceptually, Clear speech can be described as overly or extremely articulated speech, with an increased effort to distinguish between sounds, often involving slowing down the speaking rate. Unlike the Lombard reflex, "Clear" speech is the result of an active communication strategy that can vary from speaker-to-speaker. Nonetheless, a variety of such strategies have been analysed in many works and certain common characteristics have emerged (Picheny et al., 1986, 1989; Krause and Braida, 2004a,b; Drullman et al., 1994a,b; Hazan and Simpson, 1996; Hazan and Markham, 2004; Hazan and Baker, 2010, 2011; Bradlow et al., 2003; Ferguson and Kewley-Port, 2002, 2007; Liu, 2006; Amano-Kusumoto and Hosom, 2011). Specifically, compared to "casual" or "conversational" speech, Clear speech has been shown to exhibit: decreased speaking rate (with increased vowel duration and longer, more frequent pauses), increased pitch, increased consonant energy, expanded vowel space (with corresponding F1 and F2 shifts), spectral flattening or increased energy at higher frequencies and increased modulation depth in the temporal signal envelope. Of these modifications, it appears that the spectral flattening and especially vowel space expansion are among the most influential phenomena (Amano-Kusumoto and Hosom, 2011).

While Lombard and Clear speech have individually been well-studied in the literature, the styles are rarely addressed simultaneously. One exception adopting a similar approach to this work can be found in Skowronski and Harris (2006), in which commonalities between the two styles in terms of energy re-distribution between voiced and unvoiced parts of speech are exploited in developing speech intelligibility enhancement algorithms. Unlike Skowronski and Harris (2006), the present work focuses on spectral phenomena observed in Lombard and Clear speech. First, the same acoustic analyses are applied to distinct Lombard-normal and Clear-casual corpora in an effort to highlight significant properties of the styles and compare corresponding observations. As can be discerned from the discussions above, previous studies have mentioned varying degrees of spectral flattening associated with Clear speech and formant shifts for the Lombard effect. However, the relative extent of these modifications has not been compared between the two styles. While the corpora in this work are distinct, the analyses and processing are common and key observations are drawn in comparing statistics of each intelligible style with its respective counterpart, thus remaining consistent within each corpus. Consequently, analyses in the present work take initial steps towards merging studies of Lombard and Clear speech. Nonetheless, the underlying questions ultimately concern how to exploit the spectral characteristics observed in Lombard and Clear speech in designing algorithms for speech signal intelligibility enhancement.

Generally, speech modifications for intelligibility enhancement can be classified into several groups. First, there are techniques to enhancing intelligibility that exploit audio and signal properties, such as the amplitude compression scheme in Niederjohn and Grotelueschen (1976), dynamic range compression in Blesser (1969) and a method for peak-to-rms reduction in Quatieri and McAulay (1991). Second, certain speech intelligibility enhancement methods focus on speech-in-noise and exploit knowledge of the noise masker, such as the optimizations based on a speech intelligibility index in Sauert and Vary (2006) and the glimpse proportion maximization in Tang and Cooke (2011). Third, in the context of text-to-speech synthesis, adaptation approaches have been explored to increase intelligibility, as in Langner and Black (2005) and Raitio et al. (2011). Fourth, certain techniques aim to study the impact of particular acoustic

features of speech with respect to a given speech style. For example, the role of spectral modifications and fundamental frequency was examined for Lombard speech in Lu and Cooke (2009). For Clear speech, the relative intelligibility impact of several acoustic features was similarly examined in Krause and Braida (2004a,b). Moreover, several speech intelligibility enhancement approaches combine methods from above, such as the glimpse proportion maximization for HMM-based text-to-speech synthesis in Valentini-Botinhao et al. (2011). Finally, an extensive evaluation of the intelligibility of a variety of methods was recently carried out and described in Cooke et al. (2013). Emerging as the most successful modification from this challenge was the combination of Lombard-like Spectral Shaping (SS) and audio enhancement with dynamic range compression (DRC) proposed in Zorila et al. (2012). This SSDRC is used both as a starting-block and comparative standard for the modifications examined in this work.

Combining aspects of many of the above groups, the approach to intelligibility enhancement adopted here draws upon the human Lombard and Clear speaking style analyses for inspiration. At the same time, the techniques developed are speaker- and style-independent in that they can be applied generally to any speech signal, without requiring statistical learning or classification. Specifically, in combining established Lombard-inspired spectral shaping and audio-enhancement techniques proven to increase intelligibility (namely SSDRC, Zorila et al., 2012) with a novel Clear-speech inspired method for vowel space expansion, this work ultimately explores a generalized, "all-encompassing" effort for speech intelligibility improvement, albeit grounded in analyses of human speaking styles. In this respect, one motivation is to then examine if simultaneously mimicking the seemingly exclusive and distinct spectral phenomena, respectively spectral energy band boosting and vowel space expansion, compounds their respective intelligibility gains. Finally, the intelligibility impact of this overall approach and its corresponding modifications is evaluated objectively via an extended speech intelligibility index and subjectively through listening tests. Key observations and comparisons between the Lombard and Clear speech, along with their inspired modifications, highlight important aspects of the human speaking styles on their own and in relation to the proposed signal processing enhancement methods.

The structure of this article is as follows. First, the acoustic analyses of Lombard and Clear speech are presented in Section 2, specifically examining the average relative spectra followed by the vowel spaces. Distinctions between the Lombard and Clear styles are highlighted and their corresponding links to perceptual qualities and intelligibility are discussed. Following observations in Section 2, Section 3 then presents speech modifications aiming to improve intelligibility. The established spectral shaping (SS), along with dynamic range compression, is briefly described and then the novel approach to expand vowel space via frequency warping is detailed. Section 4 then presents objective and subjective evaluations that first examine the intelligibility of the different styles, followed by the impact of their respective modifications. Finally, Section 5 concludes and offers perspectives for future work.

2. Acoustic analyses of speaking styles

The following section applies the same acoustic analyses to distinct and well-established Lombard-normal (Cooke et al., 2006) and Clear-casual (Hazan and Baker, 2010, 2011) corpora. The goal is to briefly highlight significant spectral characteristics within each corpora and, more importantly, to examine differences in the observations for each intelligible style (Lombard, Clear) with respect to its counterpart (normal, casual).

2.1. Speech data

The Lombard (and normal) speech data is from the Grid corpora presented in Cooke et al. (2006) and Lu and Cooke (2008, 2009). The sentences have a simple 6-word structure (e.g., "place red in G 9 soon"), as defined in the Grid multi-talker speech corpus (Cooke et al., 2006). Each sentence was read and recorded both in quiet conditions (normal) and while the speaker listened through headphones to speech-shaped noise at a 96dB level (Lombard). The corpus recording and processing is detailed in Lu (2009). The Lombard speech corresponding to the highest noise level (i.e., Ninf96) in Lu and Cooke (2008, 2009) was selected so that the Lombard acoustic-phonetic characteristics would be most apparent. For the analyses in this work, 50 sentences per speaker, from 8 British English speakers (4 male, 4 female) are examined. The speech sampling rate is 16 kHz, downsampled from 25 kHz.

The Clear (and casual) speech data is from the read speech of the LUCID database (Hazan and Baker, 2010). Speakers were asked to read meaningful and syntactically simple sentence in two ways: (1) "casually as if talking to a friend," (2) "clearly as if talking to someone who is hearing impaired." The recording procedure and processing of the

corpus is described in more detail in Hazan and Baker (2011). In Hazan and Baker (2010) and Hazan and Baker (2011), it was shown that read clear speech is exaggerated compared to spontaneous clear speech. Thus, acoustic–phonetic differences of the read clear speech are more readily observable. For the analyses in this work, 50 distinct sentences per speaker, from 8 (Southern) British English speakers (4 male, 4 female) are examined. The speech sampling rate is 16 kHz, downsampled from 44.1 kHz.

2.2. Average relative spectra

The most significant spectral trait attributed with the intelligibility gain of Lombard speech is a boosting of spectral energy in a frequency region spanning the range of formants, sometimes referred to as a flattening of the amplitude spectrum (Summers et al., 1988; Lu and Cooke, 2009; Godoy and Stylianou, 2012). Spectral energy boosting in given frequency bands or tilt changes have also been attributed with intelligibility gains of Clear speech (Krause and Braida, 2004b; Amano-Kusumoto and Hosom, 2011). The following analyses explicitly examine the differences in spectral energy distributions of Lombard/Clear speech with respect to their normal/casual counterparts, specifically via the average relative amplitude spectra (Krause and Braida, 2004b; Godoy and Stylianou, 2012).

First, the speech signal analysis is pitch-asynchronous (using a 30 ms Hanning window and a 10 ms step) and DFT-based with an FFT length of 2048, while the spectral envelope of each frame is estimated by a "true" envelope of cepstral order 48 (Roebel and Rodet, 2005). Moreover, sentences from the respective corpora are rms-normalized such that the unmodified normal and casual speech each have maximal dynamic range for a 16 bit wav file. The Lombard and Clear speech sentences are then normalized to have the same rms energy as their normal and casual counterparts. In this way, the normalization steps adjust for overall energy differences between styles and their counterparts.

The "relative spectrum" for each speaker is then defined as the log-difference between the average Lombard/Clear and average normal/casual spectral envelopes, calculated using all frames. As in Godoy and Stylianou (2012), the DC component of this difference is removed in order to avoid a constant bias related to frame energy. This DC component would simply shift the relative spectrum up or down by a constant amount, without altering the overall curve shape. Explicitly, the relative spectrum is defined as follows. Let $S_n^Y(f)$ represent the spectral envelope for frame n of the condition Y, either Lombard or Clear. The condition X then represents either normal or casual, according to the Y counterpart. The spectral envelope $S_n^Y(f)$ is the true envelope parameterized by the discrete cepstral coefficients $\vec{c}_n^Y = c^Y(i), i = 0, \ldots, P$ and the order P is 48. The average spectral envelope for the condition Y over a total of N^Y frames is then given by

$$S^{\overline{Y}}(f) = \frac{1}{N^Y}\sum_{n=1}^{N^Y} S_n^Y(f) \tag{1}$$

The cepstral coefficients parametrizing $S^{\overline{Y}}(f)$ are represented by $\vec{c}_n^{\overline{Y}} = c^{\overline{Y}}(i), i = 0, \ldots P$. The relative ($R$) spectra, $S^R(f)$, is then defined as the spectral envelope generated from $\vec{c}^R$:

$$c^R(i) = \begin{cases} c^{\overline{Y}}(i) - c^{\overline{X}}(i) & i = 1, \ldots, P \\ 0 & i = 0 \end{cases} \tag{2}$$

where $\vec{c}_n^{\overline{X}}$ is calculated analogously to $\vec{c}_n^{\overline{Y}}$ and the zeroth cepstral coefficient is set to zero to remove a DC offset in the relative log magnitude spectrum.

The average relative spectra for the Lombard-normal and Clear-casual cases are shown in Figs. 1–4, for individual speakers as well as the overall average for the respective corpora. As expected, for the Lombard case, a consistent trend that is evident is a boosting of energy in the 500–4500 Hz region comprising the broad frequency range in which formants are located. This is observable across all of the speakers in Figs. 1 and 3, as well as on average in Fig. 2. On the other hand, for the Clear-casual relative spectra, significant variation is observed across speakers in Fig. 4, indicating that different strategies were employed. In particular, for the speakers with the most exaggerated Clear-casual differences (M11 and F14), there appears to be an actors or speakers formant (i.e., a concentration of spectral energy between 3 and 4 kHz, shown to be present for trained actors in Leino et al., 2011), which is an interesting observation that merits further investigation. Nonetheless, as is evident in Figs. 1 and 2, the overall spectral energy differences for the Clear speaking styles with respect to casual speech pale in comparison to the Lombard effect.

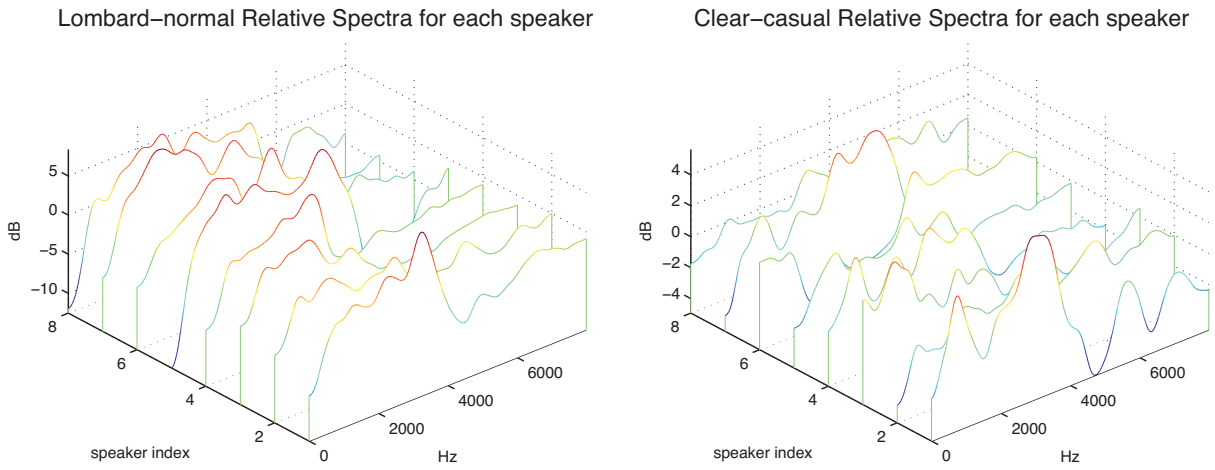

Fig. 1. Relative spectra (difference between average Lombard/Clear and normal/casual spectral envelopes) for all speakers. Lombard-normal (left) and Clear-casual (right): 1–4 are male and 5–8 are female for both corpora.

2.3. Vowel spaces

Increased vowel space (i.e., greater vowel discrimination) has been linked to higher speech intelligibility (Picheny et al., 1986; Ferguson and Kewley-Port, 2002; Hazan and Markham, 2004). In particular, for Clear speech, vowel space expansion has been shown to be one of the most evident contributors to increased speech intelligibility (Ferguson and Kewley-Port, 2002). On the other hand, vowel space expansion or corresponding formant shifts in Lombard speech are less evident (Junqua, 1993; Summers et al., 1988; Lu and Cooke, 2008; Garnier et al., 2006; Davis and Kim, 2012). The following analyses examine and compare the vowel spaces for both speaking styles.

The vowel spaces in this work have been generated in a novel way, described as follows. First, in order to isolate the vowel instances in both the Lombard (Grid) and Clear (LUCID) corpora, all of the speech was segmented using an HTK-based audio-to-text aligner. No manual corrections were performed. For each vowel instance, formant analysis is performed using the Praat algorithm exploiting the Burg estimation method (Boersma and Weenink, 2010). The representative pair of F1 and F2 values for each vowel instance is then taken as the values at the center of the speech segment. For each vowel, the mean over all of the vowel instances is trimmed, with 95% of the data kept, in order to limit the influence of potential outliers. Then, the convex hull (i.e., a polygon fit that encompasses all of the data points)

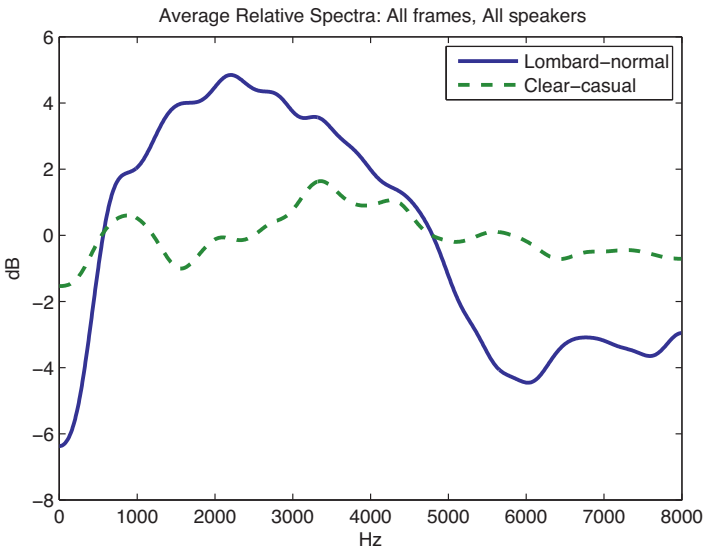

Fig. 2. Average relative spectra for all speakers and all frames in each corpus.

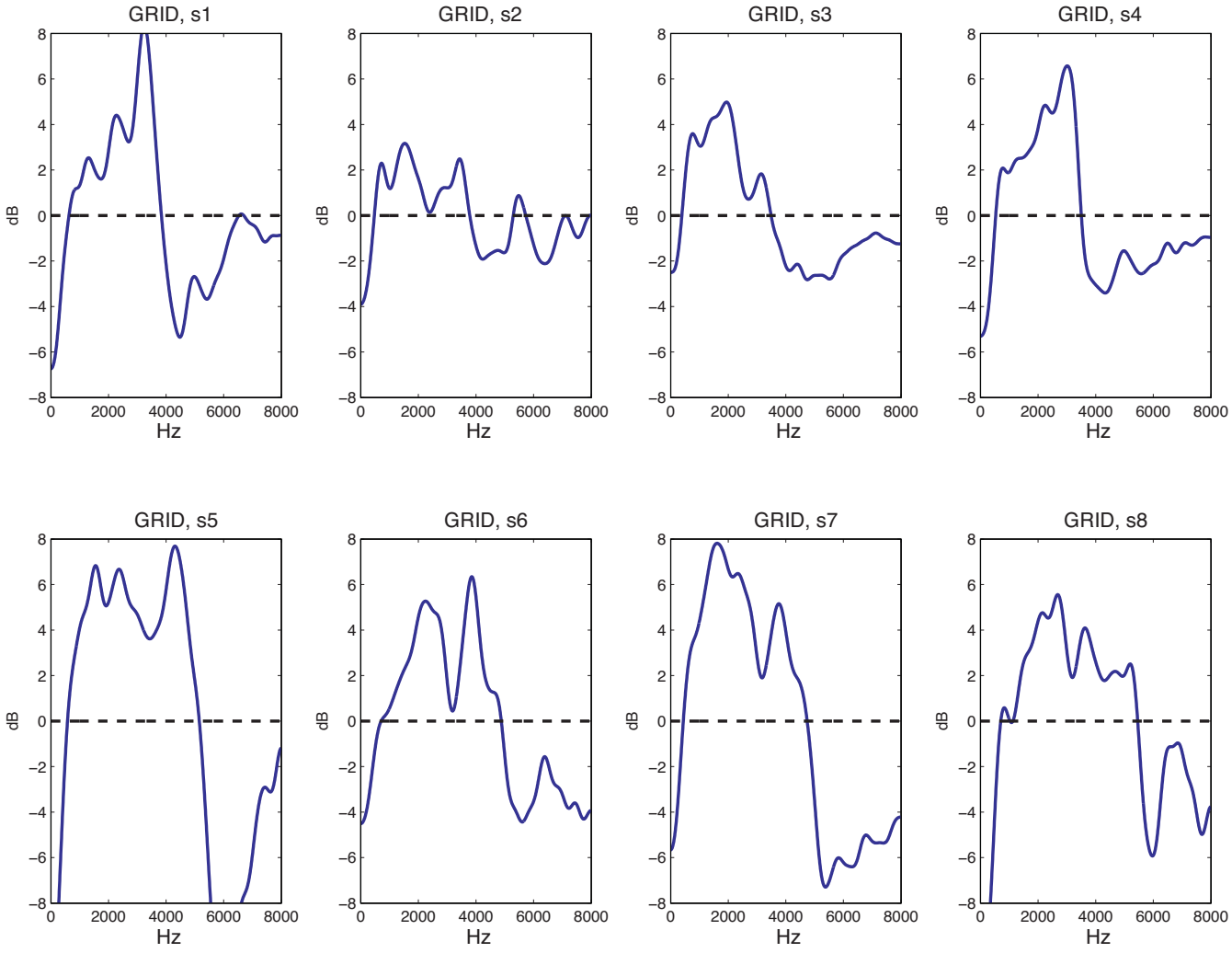

Fig. 3. Lombard-normal: relative spectra for each speaker.

is calculated in order to represent the vowel space area, as it effectively captures the maximal area that the points in the vowel space span.

Fig. 5 shows the vowel spaces calculated using all of the vowel instances for all of the speakers in the indicated corpora, while Figs. 6 and 7 show the vowel spaces for the individual speakers. Additionally, Table 1 indicates the convex hull vowel space areas for the overall average plot in Fig. 5 as well as the average of the male (M) and female (F) speakers for each style. It is apparent from Fig. 5 and Table 1 that the Clear speech vowel space area is expanded with respect to that of the casual, while this is not the case for the Lombard speech (i.e., the Lombard vowel space is not expanded). Moreover, in Fig. 7, the vowel space expansion for Clear speech is consistent across speakers, with the convex hull shape and orientation also remaining largely intact. Conversely, for the Lombard case in Fig. 6, the convex hull shape, area, and orientation are more erratic. One consistent observation for the Lombard speech, however, is an increased F1, as has been reported previously in the literature. An increased F1 is also related to the increase

Table 1
Average vowel space area ($\times 10^5$ Hz^2) determined by the convex hull, given for all speakers as well as for male (M) and female (F) speakers separately.

	Normal	Lombard	Casual	Clear
ALL	**1.78**	**1.78**	**2.32**	**3.93**
M	1.10	1.25	1.16	2.44
F	2.50	2.25	3.58	5.15

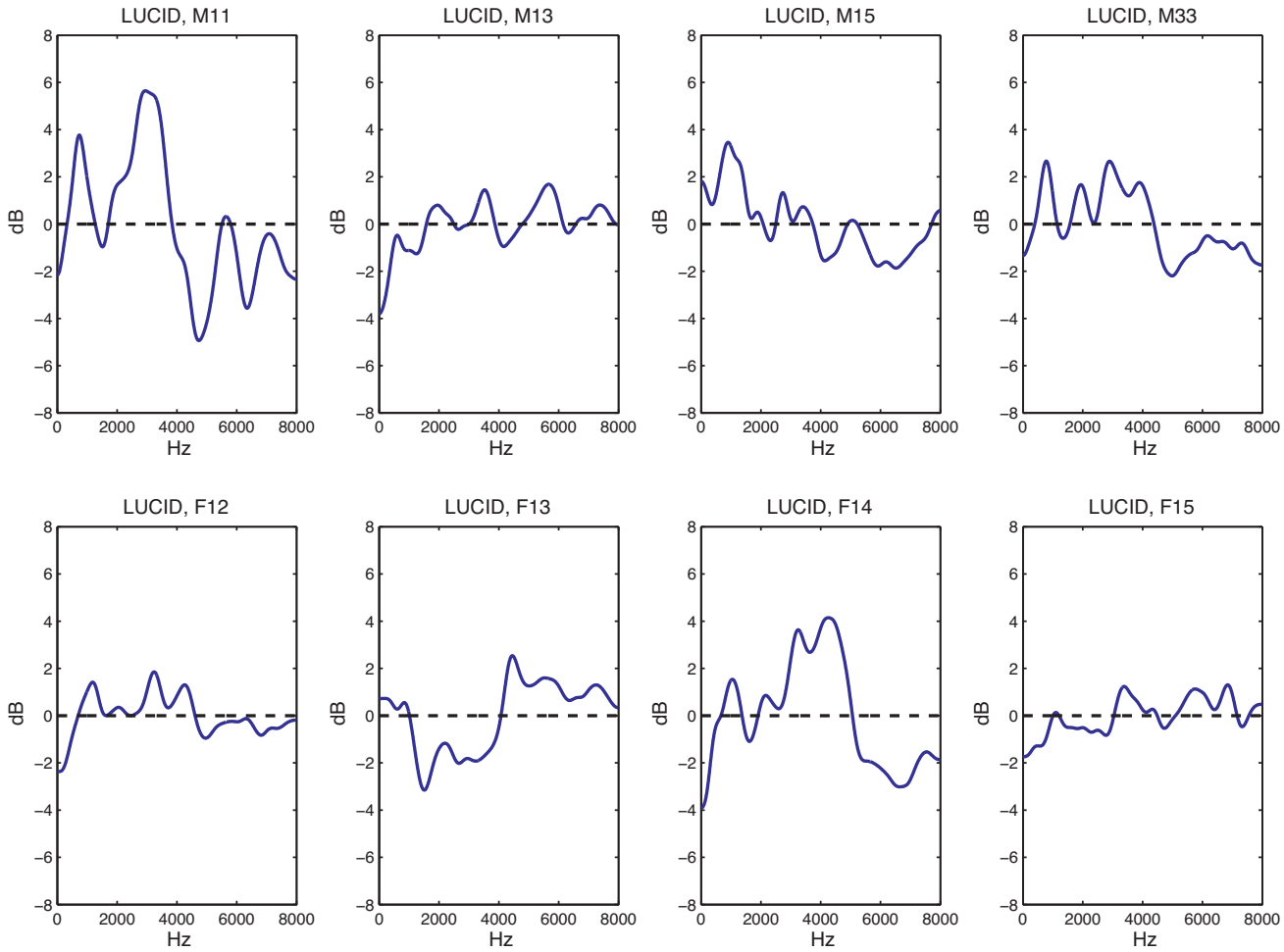

Fig. 4. Clear-casual: relative spectra for each speaker.

in fundamental frequency that is part of the Lombard effect (Lu and Cooke, 2008). However, it has been shown that increases in fundamental frequency do not yield intelligibility gains (Lu and Cooke, 2008). Ultimately, examining the trends in Fig. 5, the shifts in F1 and F2 observed for Lombard and Clear speech (with respect to their normal and casual counterparts) can be described respectively as a *translation* and *expansion* of the vowel spaces.

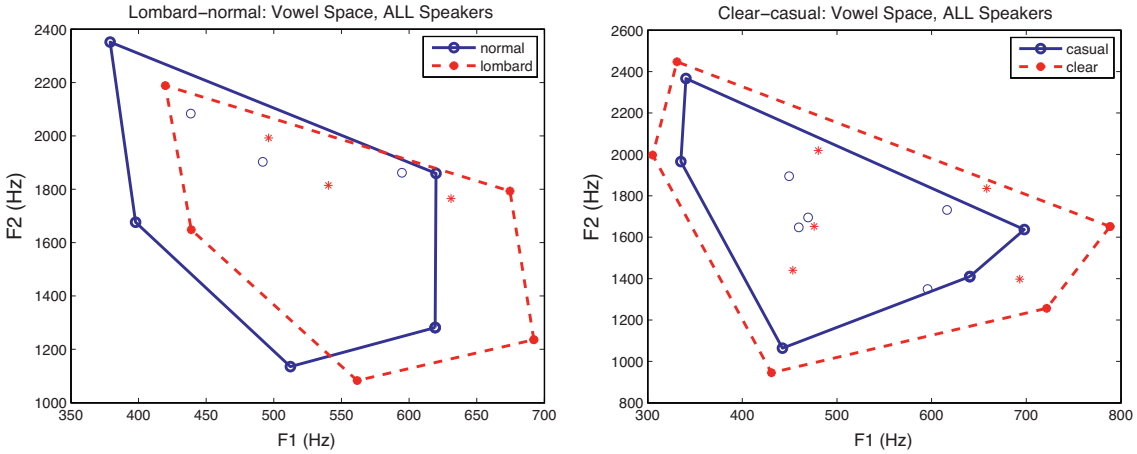

Fig. 5. Average vowel spaces: Lombard and normal speech (left), Clear and casual speech (right). The trimmed mean of F1 and F2 for each vowel is calculated across all vowel instances in each corpus. The convex hull (polygon) represents the maximal vowel space area. The two styles (Lombard and Clear) respectively produce a translation and expansion of the vowel space.

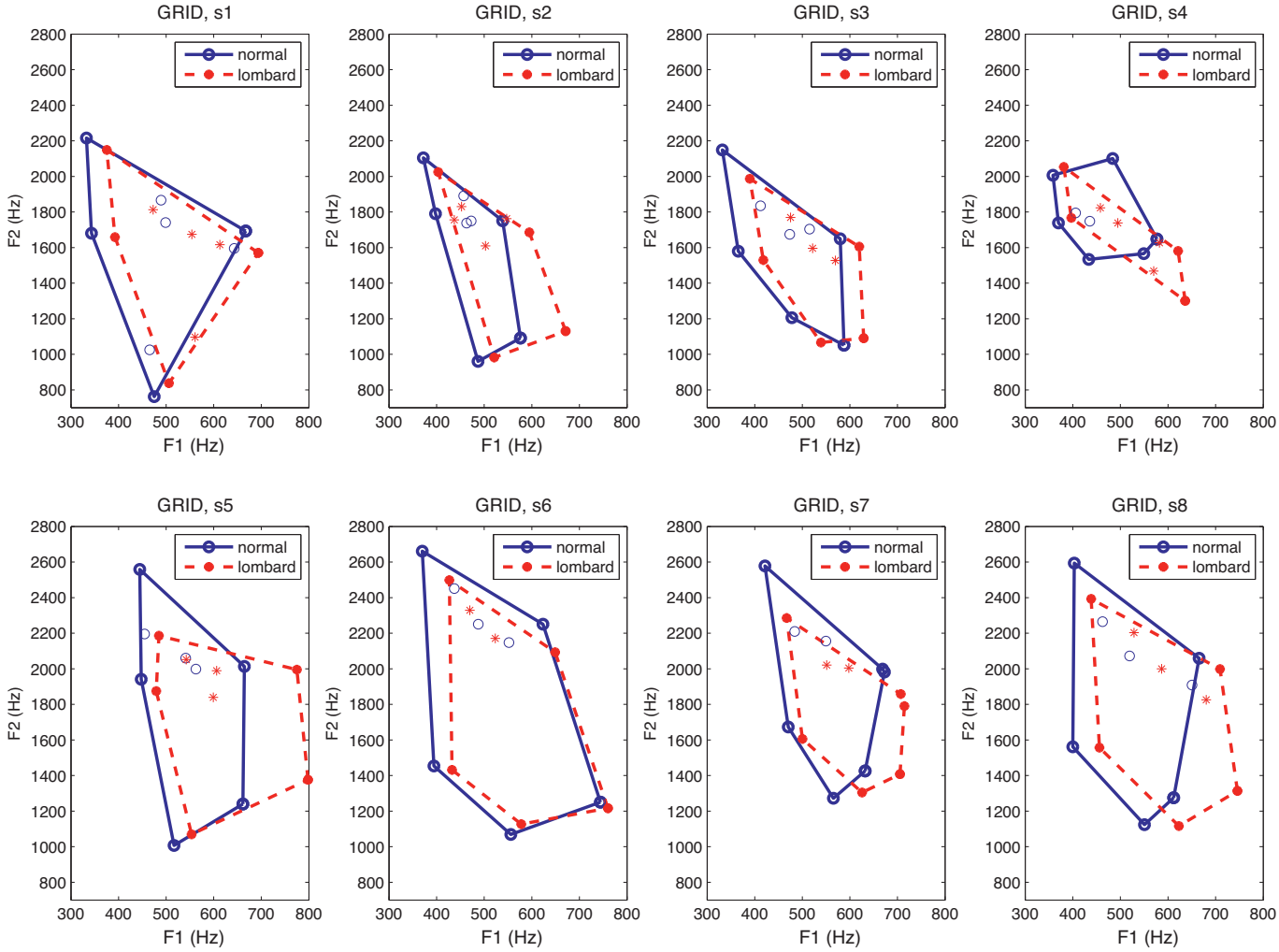

Fig. 6. Lombard-normal vowel spaces for each speaker.

2.4. *Links to perceptual traits and intelligibility*

The observations from statistics on the acoustic properties of Lombard and Clear speech highlight distinctions of the styles that are linked to their respective perceptual traits. First, considering the average relative spectra, the Lombard spectral band boosting translates into "louder" speech, as shown explicitly in Godoy and Stylianou (2012). Conversely, Clear speech does not exhibit this increase in spectral energy that augments loudness. This observation indicates that Clear speech is a more subtle phenomena with respect to re-distributing energy across spectral bands; whereas the Lombard effect can be seen as a more "brute-force" reflex to boost energy to increase audibility of important (formant) information. Then considering the vowel spaces, the expansion observed for Clear speech acoustically captures the perceived "overly" or highly articulated characteristic of the style. Whereas, for the Lombard effect, there is no indication in the vowel space analyses of increased or more deliberate articulation.

In addition to perceptual traits, the acoustic analyses of Lombard and Clear speech can be linked to respective intelligibility gains of the styles. The intelligibility advantage of the corpora examined in this work have been illustrated in Lu and Cooke (2008), Koutsogiannaki et al. (2012) and Hazan and Baker (2011), for the Lombard and Clear cases respectively. However, these studies have not examined the relative spectra and vowel spaces together, as in the present work. Consequently, re-interpreting the reported intelligibility scores for individual speakers here, provided in Tables 2 and 3, offers new insight.

Beginning with the Lombard speech, an overall increase of 25% in correct keyword identification was reported over the normal speech when SSN was added at −9 dB SNR (Lu and Cooke, 2008). More specifically, considering Table 2 and Fig. 3, there are several observations linking the relative spectra to the individual speaker intelligibility

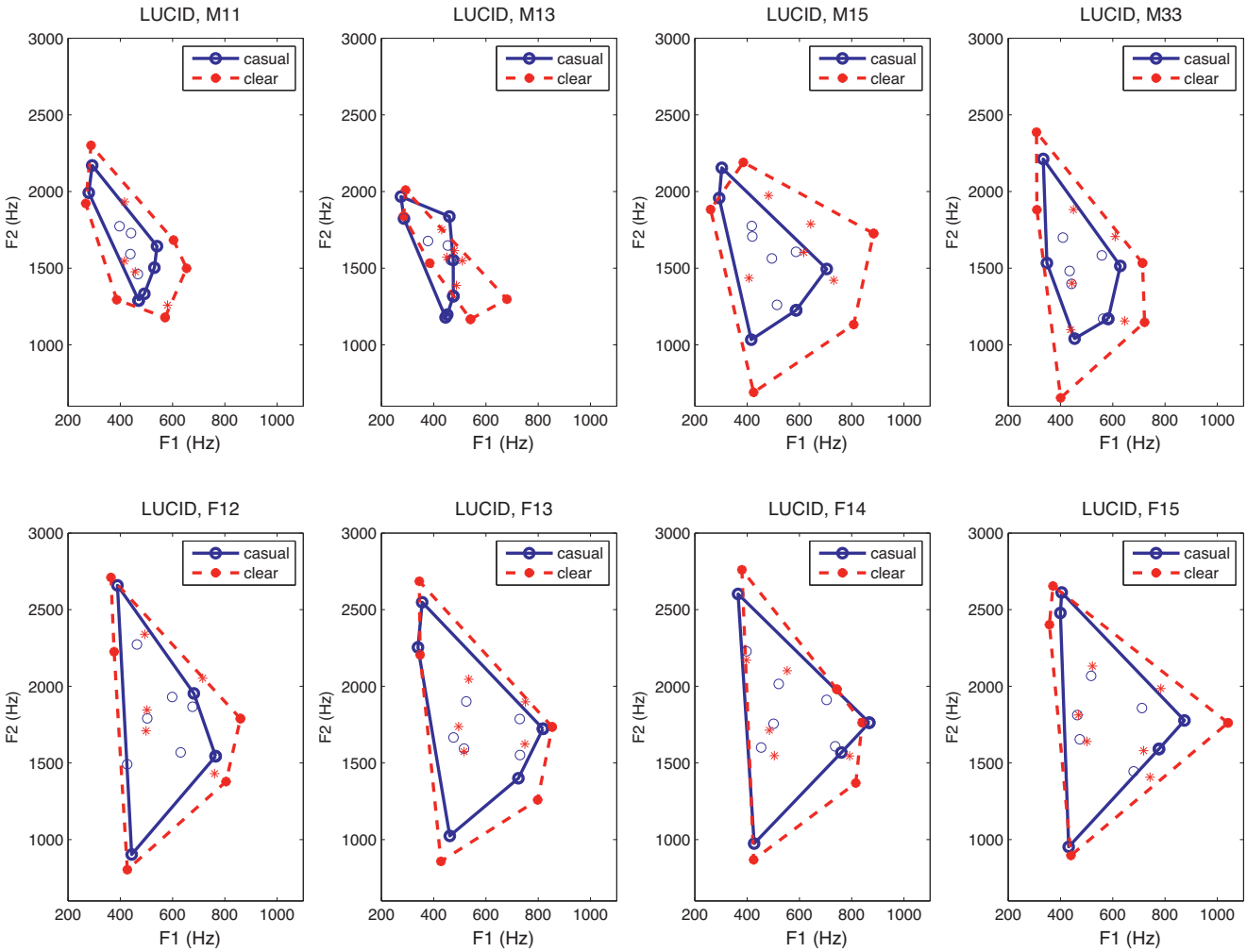

Fig. 7. Clear-casual vowel spaces for each speaker.

ratings. First, speaker 2 has the least intelligible Lombard speech and the relative spectra is the "flattest." That is, the spectral energy boosting is the smallest. Additionally, speaker 5 exhibits significant spectral energy boosting in a full 500–4500 Hz band and she is among the most intelligible speakers with a large gain in Lombard over normal speech. In addition to the average relative spectra, there are also apparent links between the speaker intelligibility ratings in Table 2 and vowel space area shown in Fig. 6. For example, speaker 1 exhibits the largest vowel space of the male speakers and has the most intelligible Lombard and normal speech. However, the gain of his Lombard over normal speech is the smallest. Moreover, of the female speakers, speaker 7 has the smallest vowel space and is the least intelligible female both for Lombard and normal speech. Finally, the above observations would indicate that the spectral energy boosting does generally impact relative gains in speaker intelligibility and speakers with a larger vowel space tend to be more intelligible.

Table 2
Intelligibility ratings (percent correct keyword identification) for the individual speakers in the Lombard Grid corpus, from the authors of (Lu and Cooke, 2008; Lu, 2009). The normal and Lombard speech correspond to the "quiet" and "Ninf96" conditions referenced in Lu and Cooke (2008) and Lu (2009). The columns correspond to the speaker and the rows in the table represent the conditions, with the last row showing the Lombard-normal gain (difference between the first two rows).

	1	2	3	4	5	6	7	8
Normal	76.9	35.0	47.3	49.6	57.7	61.5	45.0	61.9
Lombard	82.7	51.2	66.5	75.8	77.7	79.6	55.0	73.1
Gain	*5.8*	*16.2*	*19.2*	*26.2*	*20.0*	*18.1*	*10.0*	*11.2*

Table 3
Intelligibility ratings (1 – clear, 7 – unclear) for the individual speakers examined in this work from the LUCID corpus. The ratings were calculated as part of the work referenced in Hazan and Baker (2011). The columns correspond to the speaker and are ordered beginning with the most intelligible. The rating for the Clear speech (VOC condition in the Diapix task) is provided.

	F15	M15	F12	F14	F13	M33	M13	M11
Clear	1.9	2.1	2.1	2.2	2.8	2.8	3.15	3.8

Now considering Clear speech, the results in Koutsogiannaki et al. (2012) indicate an intelligibility score improvement of about 50% over casual speech in the case of SSN at a 0 dB SNR. A more detailed study of the individual speaker variability in the LUCID corpus (referenced in Hazan and Baker, 2011, p. 2146) provides the intelligibility ratings given in Table 3 for the individual speakers. Examining Table 3 and Fig. 7, these quoted ratings effectively corroborate that large vowel space and the greatest vowel space expansion yield the most intelligible speech. That is, the female speakers are more intelligible, overall, and the speakers exhibiting the greatest expansion in vowel space (F15 and M15) are the most intelligible. Furthermore, in the present work, while expanded vowel space is shown to be a prominent spectral feature of the Clear speech, differences in spectral energy are less apparent. Moreover, in comparing Fig. 4 with the quoted intelligibility ratings, these differences in spectral energy do not appear to play a significant role in determining the speaker intelligibility, unlike the case for the Lombard speech.

3. Speech modifications inspired by Lombard and Clear styles

Underlying the acoustic analyses in this work is a motivation to enact speech modifications to increase intelligibility. The approach to determining these modifications was to first study two speaking styles (i.e., Lombard and Clear) that human beings employ to increase intelligibility in different contexts, depending on whether the speaker is immersed in noise. The analyses in Section 2 indicated that there are significant, consistent acoustic modifications that are uniquely observable in these two respective styles: spectral energy boosting to increase loudness and vowel space expansion to increase articulation. Moreover, both of these phenomena are linked to gains in intelligibility. However, these key acoustic features of Lombard and Clear speech appear to be mutually exclusive in human speech production, due potentially to environmental reasons (e.g., speaker feedback) or physiological limitations.

Fortunately, signal processing algorithms are capable of enacting both of these modifications simultaneously. The following sections respectively address these two spectral phenomena, with a particular focus on the less-commonly explored vowel space expansion. That is, a Lombard-inspired fixed spectral gain filter from the SS in Zorila et al. (2012) is briefly presented and discussed first. Then, inspired by Clear speech analyses, a novel approach for vowel space expansion via frequency warping is more extensively detailed and examined.

3.1. Lombard speech-inspired spectral shaping fixed filter

In order to increase loudness and mimic Lombard speech, a corrective filter can be applied on normal speech, as demonstrated in Godoy and Stylianou (2012) and Lu and Cooke (2009). The form of this corrective filter can be speaker dependent or independent. In any case, the main shape of the filter should follow the trend observable in the form of the overall average Lombard-normal relative spectra shown in Fig. 2. Additionally, as shown in (Godoy and Stylianou, 2012) and recalled in the left plot of Fig. 8, selecting only the most objectively intelligible Lombard speech frames in defining this filter simply exaggerates the scale. Specifically, the left plot of Fig. 8 shows the average relative spectrum for the (time-aligned) Lombard and normal speech in the Grid corpus, with the dashed line representing this relative spectrum calculated using only the Lombard frames with the highest extended SII (Godoy and Stylianou, 2012) (note that this objective metric is employed later in Section 4). This plot is shown next to the fixed filter from SS ($H^r(f)$), described in Zorila et al. (2012), in Fig. 8 in order to explicitly demonstrate the shared characteristics of the filters. Ultimately, the SS fixed filter can thus be seen as a speaker-independent, simple version of a Lombard-inspired spectral gain. Moreover, this SS fixed filter has proven effective in an extensive evaluation of speech intelligibility enhancement modifications (Cooke et al., 2013). Consequently, the SS from Zorila et al. (2012) will be used in this work as the representative Lombard-inspired modification.

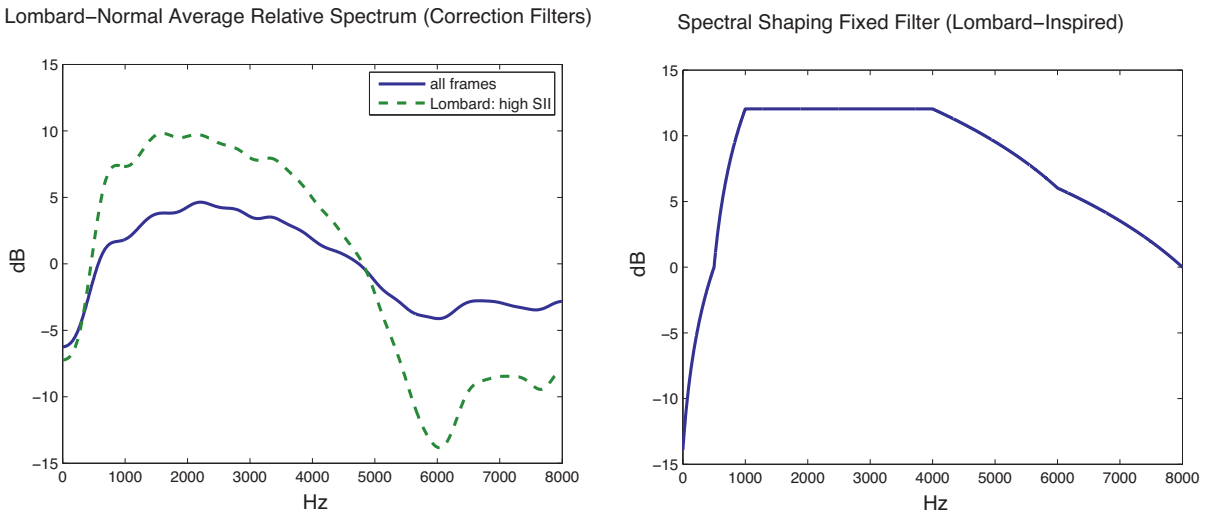

Fig. 8. The Lombard-normal relative spectrum from Godoy and Stylianou (2012) (left). The SS fixed filter $H^r(f)$ from Zorila et al. (2012) (right).

In addition to the fixed filter, the SS described in Zorila et al. (2012) and evaluated in Cooke et al. (2013) also incorporates adaptive components, including peak-sharpening ($H^s(f)$) and pre-emphasis ($H^p(f)$) filters. However, these adaptive filters are less dramatic compared to the fixed filter shown in Fig. 8 (Zorila et al., 2012). Furthermore, considering the time-domain, the dynamic range compression (DRC) paired with SS enhances intelligibility significantly by decreasing the peak-to-rms ratio of the signal energy in time, effectively increasing loudness (Blesser, 1969; Quatieri and McAulay, 1991). As the pairing of SS with DRC was shown to be highly successful at increasing intelligibility in Cooke et al. (2013), this will be used as the baseline modification in the evaluations presented in Section 4. Accordingly, one of the goals is then be to examine if incorporating a Clear speech-inspired frequency warping method for vowel space expansion into SSDRC could further increase intelligibility.

3.2. *Clear speech-inspired frequency warping for vowel space expansion*

Unlike the Lombard-inspired spectral gain filtering, methods for vowel space expansion are rarely explored, as reliably manipulating formants is a challenging feat discouraged by limitations in accurate formant estimation and treatment. That said, general upwards shifts of formants (McLoughlin and Chance, 1997; Valentini-Botinhao et al., 2011) and statistical methods for spectral envelope transformation to mirror that of Clear speech, including the vowel space expansion, Mohammadi et al. (2012) have been explored. However, these techniques do not explicitly isolate and evaluate the intelligibility gains of vowel space expansion. The following work focuses especially on this task, achieving vowel space expansion via frequency warping inspired by Clear speech analyses.

Typically used in voice conversion (Valbret et al., 1992; Godoy et al., 2012; Erro et al., 2010) or speech recognition (e.g., VTLN), frequency warping is employed here in a novel manner as a means for vowel space expansion (Godoy et al., 2013). The appeal of frequency warping for this expansion is that it offers a way of shifting speaker formants, while both avoiding notable speech degradations and limiting dependence on accurate formant detection. The proposed frequency warping approach for vowel space expansion can be described in two stages.

First, inspired by the formant shifts observed in the Clear speech vowel expansion, a curve of generalized warping shifts $\Delta(f)$ is established. Fig. 9 shows these observed Clear speech formant shifts (i.e., $F1_{Clear} - F1_{casual}$ and $F2_{Clear} - F2_{casual}$) alongside the inspired generalized warping shifts. The practical considerations taken into account when generating $\Delta(f)$ (via several trials) included exaggerating the observed formant shifts to overcome harmonic separation and adjusting slope to avoid F1 overlapping with F2 (Godoy et al., 2013). Overall, it is evident from Fig. 9 that the resulting shape of the generalized warping curve mirrors that observed in the casual-to-Clear F1 and F2 shifts, especially in comparison with the piecewise polynomial fit shown with solid lines.

The second stage of the proposed frequency warping algorithm then involves sampling the $\Delta(f)$ curve, using detected spectral peaks, on a frame-by-frame basis. More explicitly, the frequency warping takes the form of a filter that is applied to the DFT amplitude spectrum of each frame. Note that, when combined with SSDRC, the frequency warping is thus

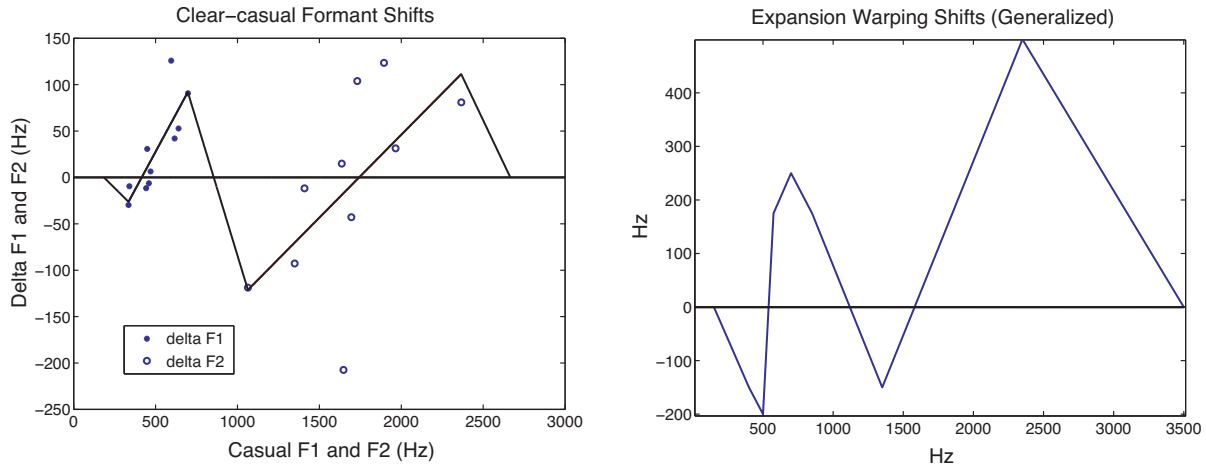

Fig. 9. Observed formant shifts from casual to Clear speech (left). $\Delta(f)$ – generalized curve of exaggerated warping shifts (right).

an extra filter incorporated in the spectral shaping, as is shown explicitly later. Now, recall that the spectral envelope from frame n of unmodified speech is $S_n^X(f)$. The frequency warping filter for frame n, $H_n^W(f)$ is then defined as

$$H_n^W(f) = \frac{S_n^W(f)}{S_n^X(f)} \tag{3}$$

where $S_n^W(f)$ is the warped spectral envelope

$$S_n^W(f) = S_n^X(W_n^{-1}(f)) \tag{4}$$

and $W_n(f)$ is the warping function for frame n. The form of the warping function is piecewise linear, as in Godoy et al. (2012) and Erro et al. (2010), and is given for $f \in [f_{n,i}^X, f_{n,i+1}^X]$ by

$$W_n(f) = A_{n,i} f + B_{n,i} \tag{5}$$

where $f_{n,0}^X = f_{n,0}^W = 150$, $f_{n,M_n+1}^X = f_{n,M_n+1}^W = 3500$, and

$$A_{n,i} = \frac{f_{n,i+1}^W - f_{n,i}^W}{f_{n,i+1}^X - f_{n,i}^X} = \Delta(f_{n,i+1}^X) - \Delta(f_{n,i}^X) \tag{6}$$

$$B_{n,i} = f_{n,i}^W - A_{n,i} f_{n,i}^X \tag{7}$$

$$f_{n,i}^W = f_{n,i}^X + \Delta(f_{n,i}^X) \tag{8}$$

and $f_{n,i}^X$ indicates the frequency of the ith spectral envelope peak detected in frame n, $i = 1, \ldots M_n$. Note that the peaks f_n^X are detected as maxima (following a minima lower by at least 10%) of the tilt-normalized (the first two cepstral coefficients are removed) spectral envelope. With $W_n(f)$ defined in this way, the warping filter $H_n^W(f)$ is calculated and applied to the amplitude spectrum $E(f)$ of each frame, ensuring that vowels spaces are expanded, without need for speech segmentation and labelling. Furthermore, it should be emphasized that the focus on overall average trends for the vowel space expansion makes the proposed algorithm speaker-independent and thus generalized, as is shown in the next subsection.

3.2.1. Results of frequency warping on vowel spaces

Fig. 10 shows the vowel space expansion resulting from the proposed method being applied to *both* casual speech and normal speech (from the Lombard corpus). The warped vowel space is generated in the same way as the normal/casual and Clear/Lombard vowel spaces in Section 2.3, but with the data being the warped normal/casual sentences. That is, the frequency warping is applied on the casual/normal speech and the resulting speech is analysed in the same way as described in Section 2.3. Fig. 10 shows that both the casual and normal speech vowel spaces are successfully expanded. The warped-casual and warped-normal vowel space areas (3.58 compared to 2.32×10^5 Hz^2 and 1.78 compared to

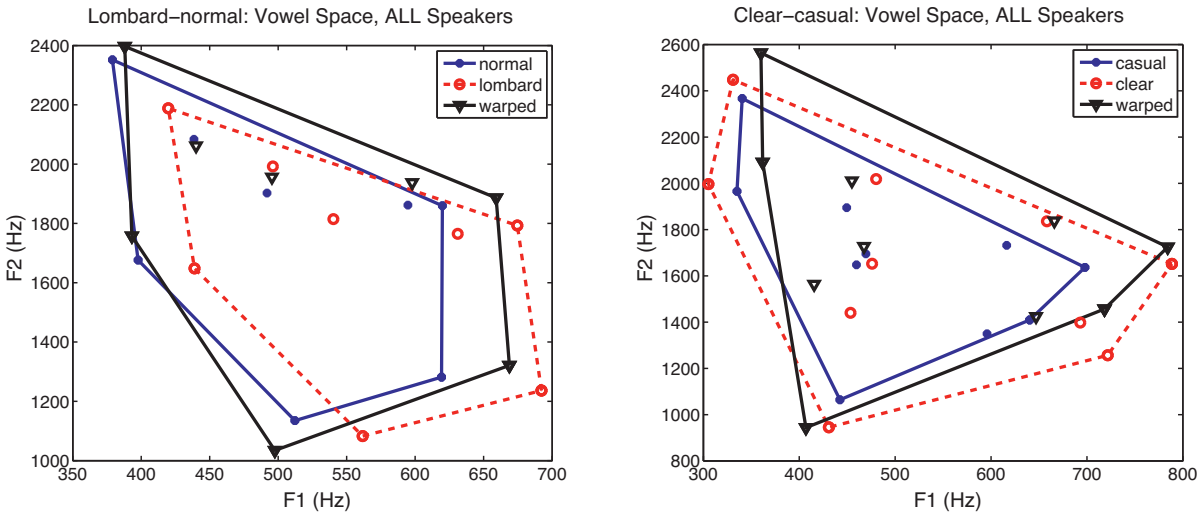

Fig. 10. (Left) Lombard, normal and normal-warped vowel spaces, with respective areas: 1.78, 1.78 and 2.38 ($\times 10^5$ Hz^2). (Right) Clear, casual and casual-warped vowel spaces, with respective areas: 3.93, 2.32 and 3.58 ($\times 10^5$ Hz^2).

2.38×10^5 Hz^2 respectively), also confirms this expansion emulating that observed in Clear speech. Moreover, the structure of the vowel spaces are largely maintained for both corpora, ensuring that the perceptual distinctions between vowels is respected, with only the distance or discriminability between them being increased. It should be noted that the proposed frequency warping approach is a generalized approximation, rather than deterministic replication, of the vowel space expansion observed in Clear speech. In this way, the vowel space expansion is achieved on both the casual and normal speech, without explicit vowel or formant identification.

4. Evaluations

The evaluations in this section serve several purposes. First, comparisons are made on the intelligibility of the Lombard and Clear speech corpora for different SNR (high and low). Second, the intelligibility impact of the proposed frequency warping for vowel space expansion is examined over unmodified normal and casual speech, in an effort to consequently isolate any potential intelligibility gain associated with vowel space expansion. Finally, the all-encompassing modification combining the Lombard-inspired SS with the Clear-speech inspired frequency warping, in addition to the proven DRC, is evaluated in order to examine if the addition of frequency warping could further enhance the intelligibility gain of SSDRC. Moreover, these modifications are carried out on the normal and casual speech, thus attesting to their generalizability (i.e., indicating that they are not speaker or corpus dependent).

Before the evaluations, the following briefly described the modified speech synthesis. Considering the full combination of the SS and frequency warping filters, the modification for frame n, with amplitude spectrum $E_n(f)$, is given by

$$\hat{E}_n(f) = E_n(f)H_n^p(f)H_n^s(f)H^r(f)H_n^W(f) \tag{9}$$

That is, in the case of SSDRC with frequency warping, all of the filters in Eq. (9) are involved, while only the first three or last are considered respectively with only SSDRC or only frequency warping. Thus, in modification, the appropriate filters are applied to each speech frame, then synthesis using the original phases from analysis and overlap add (OLA) with the same pitch asynchronous analysis time instants. As a last step, DRC is applied (when specified) to the modified sentence. Finally, all modified sentences are rms-normalized to match the energy of the original, unmodified sentence.

4.1. Objective evaluations with the speech intelligibility index

In order to objectively evaluate the intelligibility of the Lombard and Clear styles, the extended SII is examined (Rhebergen et al., 2006), as in Godoy and Stylianou (2012) and Zorila et al. (2012). Histograms of the extSII values for the different speaking styles for all speech frames (normalized to sum to one) are shown in Fig. 11 along with the

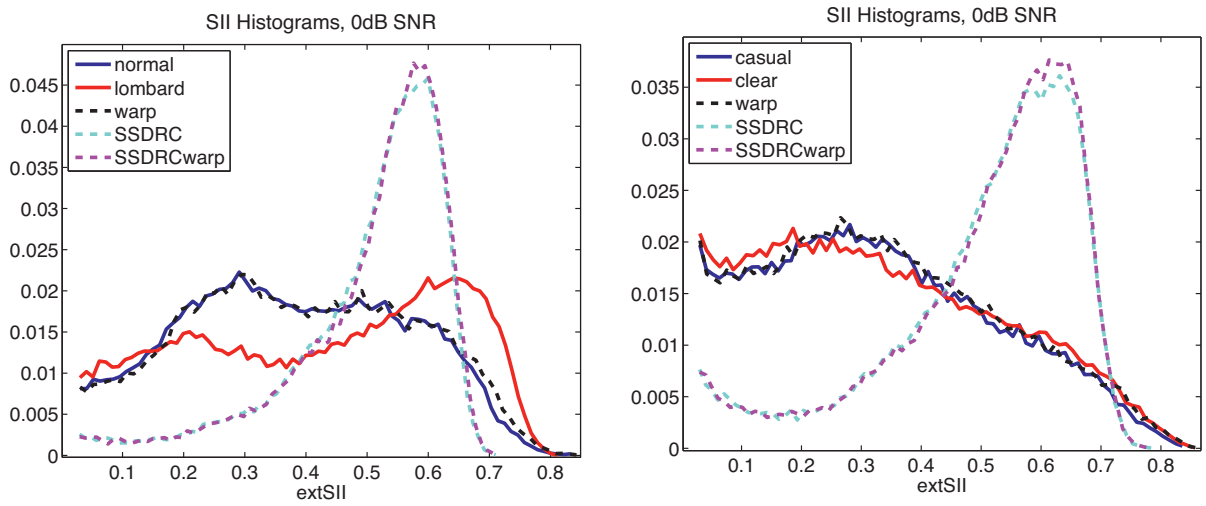

Fig. 11. Histograms of extended SII with SSN added to yield 0 dB SNR. Lombard and normal, with modifications (left), Clear and casual, with modifications (right). The curves from modified speech are shown with dashed lines.

Table 4
Average extSII for the distributions shown in Fig. 11. "Style" corresponds to Clear/Lombard speech accordingly.

	Orig.	Warp	SSDRC	Both	*Style*
Casual	.311	.316	.542	.547	*.312*
Normal	.374	.383	.538	.543	*.454*

average SII of the distributions in Table 4. That is, the calculated extSII for each frame is accumulated into a histogram, which is then normalized by the total number of frames. The extSII is calculated using speech shaped noise (SSN) added in order to yield a 0 dB signal-to-noise ratio (SNR).

One key observation from Fig. 11 is that, while Lombard speech is objectively more intelligible than normal speech, Clear speech is not shown here with the extSII to be more objectively intelligible than casual. Considering the modifications, with spectral shaping mimicking the Lombard speech, SSDRC significantly increases the objective intelligibility scores. On the other hand, the frequency warping for vowel space expansion does not demonstrate an objective gain with the extSII (neither over unmodified speech, nor over SSDRC).

The above observations on the extSII highlight an important limitation in objectively evaluating intelligibility. It has been well-established that Clear speech is indeed more intelligible than casual speech (Picheny et al., 1986; Krause and Braida, 2004b; Hazan and Baker, 2010; Bradlow et al., 2003; Ferguson and Kewley-Port, 2002, 2007; Liu, 2006), however, the extSII examined here is unable to capture this intelligibility gain. Following from the previous analyses, it was observed that Clear speech incorporates more subtle spectral modifications (e.g., vowel space expansion) than the Lombard speech, which boosts spectral energy to increase loudness and audibility of formants. Unfortunately, these more subtle features of Clear speech are evidently not valorised in the objective intelligibility evaluation here with the extSII. Consequently, listening tests are required to more thoroughly capture the intelligibility gains of the different styles and modifications.

4.2. *Subjective evaluations via listening tests*

In formal listening tests,[1] 20 native English-speakers heard Grid and LUCID sentences, with SSN added at two levels to yield 0 ("high") and −4 ("low") dB SNR, respectively. The original normal, casual, Clear and Lombard sentences were included in the test, as well as the normal and casual sentences modified using the frequency warping, SSDRC and both. In should be mentioned that, perceptually, no significant artefacts resulted from the modifications. However, voice quality was noticeably altered by both the frequency warping and SSDRC, on their own and together.

[1] Thank you to Catherine Mayo and CSTR at the University of Edinburgh for their help administering the listening tests.

Table 5
Overall percent of correct keyword identification for the evaluated conditions. "Style" corresponds to Clear/Lombard speech accordingly.

	Orig.	Warp	SSDRC	Both	*Style*
Casual, 0 dB	50	42	86	82	*79*
Casual, −4 dB	15	14	61	49	*46*
Normal, 0 dB	62	50	81	71	*67*
Normal, −4 dB	33	31	59	47	*50*

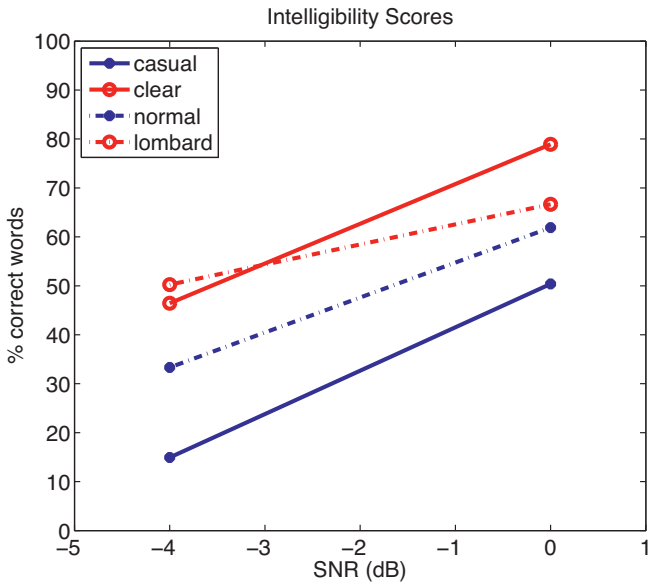

Fig. 12. Intelligibility test scores: Clear/Lombard, casual/normal.

For each listener, the test was split into two parts, involving the Grid and LUCID sentences, respectively. The order of these parts was randomized. Additionally, the sentences within each part were randomly ordered and the test was designed with the goal that speakers and conditions be equally represented (i.e., appear approximately the same number of times across the tests). When taking the test, listeners were asked to type what they think they heard after hearing each sentence once. Of the listeners, 6 were removed due to inconsistencies in their scores, using established conditions (e.g., Clear and Lombard speech) as a reference. It should be noted that the variability among listeners was quite high. Nonetheless, the average trends in intelligibility scores are shown in Table 5 and Figs. 12 and 13. In order to evaluate the statistical significance of these results, ANOVA tests were performed. Specifically, the ANOVA null hypothesis (i.e., the average values of the intelligibility scores for every condition are equal) was rejected using the F-test (SNR−4dB: $F(4, 65) > 28.719$, $p < 0.05$ SNR0dB: $F(4, 65) > 25, 307$, $p < 0.05$). Then, pairwise comparisons of the averages were performed using Fisher's least significant difference (LSD) test, with a confidence interval of 95%, in order to derive which of the groups differ significantly. The standardized difference between condition pairs is provided in Table 6. The following discussion of the listening test results is split into two parts, first focusing on comparisons between the Lombard and Clear corpora and then focusing on evaluating the intelligibility impact of the modifications.

Table 6
Results of significant difference analysis between conditions for the Grid (G) and LUCID (L) corpora. The standardized difference is given for the pairing between: Unmodified-U, Style-I, Warped-W, SSDRC-S, Both-B. Significant differences are in bold.

	U-W	U-I	W-I	U-S	W-S	U-B	W-B	S-I	B-I	S-B
G, 0	1.77	.709	**2.48**	**2.91**	**4.68**	1.42	3.12	**2.20**	**.709**	1.49
G, −4	.228	**2.03**	**2.26**	**3.02**	**3.24**	1.65	1.88	.982	.384	1.37
L, 0	1.41	**5.14**	**6.55**	**6.28**	**7.70**	**5.62**	**7.04**	1.15	.488	.657
L, −4	.041	**5.66**	**5.70**	**8.26**	**8.30**	**6.05**	**6.09**	**2.60**	.394	**2.21**

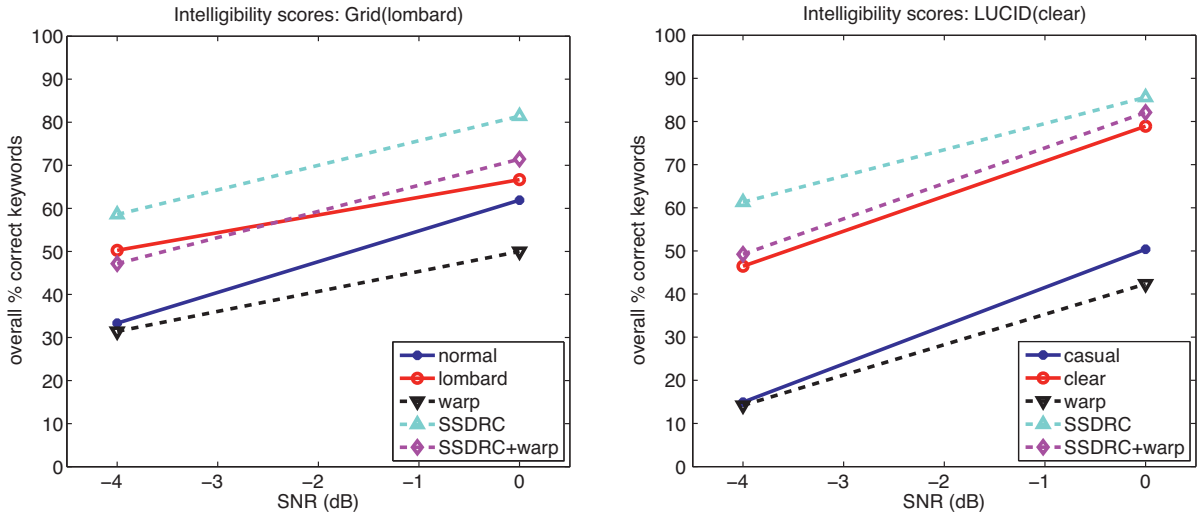

Fig. 13. Intelligibility test scores: Clear/Lombard, casual/normal, Warp, SSDRC, Both (SSDRC + Warp).

4.2.1. *Comparing the Lombard and Clear speech corpora*

Fig. 12 shows the intelligibility scores (i.e., percent of keywords correctly identified) from the listeners for each condition, at the two SNR levels. The first and last columns in Table 5 similarly indicate the percent of overall correct keywords identified for these conditions. It should be noted that the all of the conditions shown in Fig. 12 were found to be statistically significant, as shown in Table 6.

In examining Fig. 12 and Table 5, several trends are observed for the Lombard and Clear speech corpora. First, comparing the Clear and casual speech scores, there is a +31% and +29% gain in intelligibility of Clear speech over casual for low and high SNR, respectively. Thus, the intelligibility gain of Clear speech over casual is consistent across both SNR. On the other hand, in comparing the gain of Lombard over normal speech, +17% and +5% are observed for low and high SNR, respectively. Thus, the intelligibility advantage of Lombard speech over normal speech is noticeably reduced for high SNR, e.g., situations in which the noise level is low. This observation is in line with related work in Lecumberri (2012) suggesting that Lombard speech does not necessarily provide an intelligibility advantage at high SNR. Overall, at low SNR, the Lombard speech was judged to be most intelligible of the original (unmodified) speech, while at high SNR, it was the Clear speech. Considering the acoustic phenomena underlying these styles, the intelligibility scores would thus suggest that increasing articulation (as observed in the expanded vowel space) is always helping to make speech more intelligible, however, boosting audibility or loudness (as observed as increased spectral energy in formant bands) is only advantageous when hearing the speech in noisy conditions.

Considering the "standard" conditions, the normal speech from Grid was judged to be more intelligible than the casual speech from LUCID. One likely explanation for this intelligibility difference is the sentence structure used in the corpora. Specifically, the Grid corpus has a constrained pattern and identifiable structure. Consequently, it is easier to guess the keywords once some highly intelligible speech examples have been heard. For example, the structure of the Grid sentences "Bin blue at G 6 again" and "Place red by Z six now" are more similar than the LUCID sentence examples "Wasps and bees are part of summer" and "Jonathon gave his wife a bush." Moreover, the British accents of the speakers in the different corpora are different and this could also play a role in the intelligibility scores. Given these observations, the results comparing the different corpora are accordingly tempered. Nonetheless, the results of the subjective tests are consistent with observations made in related works, in particular concerning the intelligibility gains of Lombard and Clear speech at low and high SNR. In the following subsection, however, the focus on the intelligibility advantage of modified speech is on comparisons and relative gains within corpora and thus, wholly consistent in that they are isolated from differences in corpus structure and speaker accents.

4.2.2. *Evaluating the intelligibility impact of speech modifications*

Fig. 13 and Table 5 show the intelligibility scores of all conditions evaluated in the listening tests and Table 6 provides the statistical significance between the conditions. These results illustrate several trends, particularly indicating the effectiveness of the different modifications. First, the frequency warping is not shown to increase intelligibility. That

Table 7
MOS results evaluating the quality of the modified speech.

	Unmodified	Warp	SSDRC	Both
MOS	4.56	3.92	3.26	2.73

is, no gain is observed over unmodified speech, nor over SSDRC in the case of Both. However, the slight intelligibility decrease observed in these comparisons is not statistically significant in general. Nonetheless, the frequency warping, while shown to achieve vowel space expansion (emulating that observed in Clear speech), does not offer an intelligibility advantage.

Overall, SSDRC was shown to consistently yield the highest scores, for both the Grid and LUCID corpora. For both SNR, SSDRC outperformed Lombard and Clear speech (though in the latter case, the scores were comparable for high SNR). This observation on SSDRC also follows from the extSII trends observed in Section 4.1, indicating that the objective metric is effectively capturing the intelligibility gains resulting from increased loudness, as indicated explicitly in (Godoy and Stylianou, 2012). Additionally, similarly to Lombard speech, the gain of SSDRC over unmodified speech was larger for low SNR: +46% vs. +36% (casual) and +26%, +19% (normal). Consequently, the increased loudness is more effective for speech intelligibility at low SNR.

In order to begin to understand the intelligibility difference observed between the modifications and, in particular, why the proposed warping decreased the gain, listening tests to evaluate the *quality* of the modified speech were carried out. Using the casual speech as a baseline, 15 listeners were asked to rate the quality on a scale from 1-bad to 5-excellent, of the original (unmodified) speech, along with randomized sample sentences of the Warp, SSDRC and Both modifications. The MOS results are provided in Table 7. From Table 7, it is evident that the warping degrades quality by over 0.5 points, both in the casual vs. warp and SSDRC vs. Both cases. This degradation in speech quality could explain the observed reduction in intelligibility. However, Table 7 also indicates that SSDRC reduces the perceived speech quality by 1.3 points, though the intelligibility gain is large. Though the intent of SSDRC is for use in noisy conditions, evaluating the quality of the modified speech offers interesting insight and ultimately raises questions about speech naturalness versus intelligibility, a debate that will continue in future works.

4.3. Discussion

To summarize the main results of evaluations, when comparing Lombard and Clear speech, both styles are highly intelligible, albeit due to different acoustic phenomena. In particular, this work highlights the spectral energy boosting increasing loudness of Lombard speech and the vowel space expansion resulting from more deliberate articulation in Clear speech. While Clear speech is highly intelligible at both low and high SNR, the Lombard gain is more significant in the presence of higher noise levels.

When comparing the effectiveness of the modifications, the results indicate that augmenting loudness with SS and DRC is highly effective at increasing intelligibility to levels even greater than Lombard and Clear speech, with a slightly reduced gain at high SNR, much like SS's Lombard fixed-filter inspiration. On the other hand, though the frequency warping was shown to expand vowel space in Section 3.2.1, there was no gain in intelligibility (and even a slight degradation overall). The Clear speech that served as the inspiration, however, yielded significant gains in intelligibility.

The observations on the modifications and their respective inspired speaking styles highlight an important point. Though the modifications examined in this work mimic acoustic phenomena observed in human speaking styles, they are not "natural." That is, they exploit certain simple signal processing techniques (e.g., a fixed filter and frequency warping) and do not account for many of the complex features involved in human speech production. On the one hand, the simple techniques offer advantages over human speaking styles, as is the case of SSDRC. On the other hand, the modifications can also be overlooking important features that impact speech intelligibility, as is the case for the frequency warping. That is, the analyses in this work established a link between the observed vowel space expansion in Clear speech and gains in intelligibility. However, the manner in which vowel space expansion is achieved via frequency warping is not effective at increasing intelligibility, suggesting that some important (currently elusive) features are not appropriately addressed. In the end, the reasons for which the frequency warping did not yield notable gains in intelligibility is the subject of future work.

One potential explanation for the lack of effectiveness of the proposed frequency warping is that the analyses and modification focus on overall average trends. It is possible that this overall average level is not appropriate. That is to say, the level of analyses and especially modification might need to be more localized, to within vowel phonemes or even instances, for example. Additionally, the dynamic role of features (e.g., formants) in time might be more significant than the overall average values. These different levels of analyses and modifications are the subject of future work and seem apt in seeking to further isolate the acoustic phenomena responsible for the intelligibility gain of the speaking styles, especially for Clear speech.

5. Conclusions

In the present work, acoustic analyses were conducted for Lombard and Clear speaking styles in order to isolate and compare pertinent spectral phenomena that later inspire speech modifications to increase intelligibility. While Lombard speech consistently exhibits spectral energy boosting in an inclusive formant region, effectively increasing loudness, Clear speech consistently yields expanded vowel spaces, increasing vowel discrimination. Moreover, these two acoustic phenomena are shown to be largely mutually exclusive, potentially due to environmental or physiological constraints. Nonetheless, both phenomena are attributed with and shown to be linked to gains in intelligibility. Moving beyond the apparent limitations observed in the human modifications, signal processing approaches were then proposed that can simultaneously accomplish the highlighted acoustic feats of both styles. The fixed filter from SS captures the Lombard spectral boosting, while a novel Clear speech-inspired method is presented that achieves vowel space expansion via frequency warping. The effectiveness of these human-inspired modifications at increasing intelligibility was then evaluated, both comparing the methods and the corresponding speaking styles. While SSDRC was again proven effective, especially at low SNR, the proposed frequency warping did not increase intelligibility, even though the vowel spaces were expanded. Nonetheless, evaluations of the Lombard and Clear styles along with their inspired modifications provided useful insight into the effectiveness of the examined acoustic modifications and the levels of their analysis and application.

Acknowledgements

The authors thank EU Future and Emerging Technology (FET-OPEN) Project LISTA (The Listening Talker) for support and encouragement. Additionally, many thanks to Paul Iverson, Mark Wibrow, Jos Joaqun Atria and Valerie Hazan for providing the authors with the HTK aligner.

References

Amano-Kusumoto, A., Hosom, J., 2011. A review of research on speech intelligibility and correlations with acoustic features. Center for Spoken Language Understanding, Oregon Health and Science University (Technical Report CSLU-011-001).

Blesser, B., 1969. Audio dynamic range compression for minimum perceived distortion. IEEE Trans. Audio Acoust., 17.

Boersma, P., Weenink, D., 2010. Praat: Doing Phonetics by Computer.

Bradlow, A.R., Kraus, N., Hayes, E., 2003. Speaking clearly for learning-impaired children: sentence perception in noise. J. Speech Lang. Hear. Res. 46, 80–97.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition (l). J. Acoust. Soc. Am. 120, 2421–2424 (Letters to the Editor).

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., Tang, Y., 2013. Evaluating the intelligibility benefit of speech modifications in known noise conditions. Speech Commun..

Davis, C., Kim, J., 2012. Is speech produced in noise more distinct and/or consistent? Speech Sci. Technol., 46–49.

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

Drullman, R., Festen, J.M., Plomp, R., 1994a. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am. 95, 2670–2680.

Drullman, R., Festen, J.M., Plomp, R., 1994b. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Am. 95, 1053–1064.

Erro, D., Moreno, A., Bonafonte, A., 2010. Voice conversion based on weighted frequency warping. IEEE Trans. Audio Speech Lang. Process. 18, 922–931.

Ferguson, S.H., Kewley-Port, D., 2002. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 112, 259–271.

Ferguson, S.H., Kewley-Port, D., 2007. Talker differences in clear and conversational speech: acoustic characteristics of vowels. J. Speech Lang. Hear. Res. 50, 1241–1255.

Garnier, M., Bailly, L., Dohen, M., Welby, P., Loevenbruck, H., 2006. An acoustic and articulatory study of Lombard speech: global effects on the utterance. In: Interspeech.
Godoy, E., Koutsogiannaki, M., Stylianou, Y., 2013. Assessing the intelligibility impact of vowel space expansion via clear speech-inspired frequency warping. In: Interspeech 2013, Lyon, France.
Godoy, E., Rosec, O., Chonavel, T., 2012. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. IEEE Trans. Audio Speech Lang. Process. 20, 1313–1323.
Godoy, E., Stylianou, Y., 2012. Unsupervised acoustic analyses of normal and Lombard speech, with spectral envelope transformation to improve intelligibility. In: Interspeech 2012, Portland, Oregon, USA.
Hanley, T.D., Steer, M.D., 1949. Effect of level of distracting noise upon speaking rate, duration, and intensity. J. Speech Hear. Disord. 14, 363–368.
Hazan, V., Baker, R., 2010. Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style? DiSS-LPSS, 7–10.
Hazan, V., Baker, R., 2011. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. J. Acoust. Soc. Am. 130, 2139–2152.
Hazan, V., Markham, D., 2004. Acoustic-phonetic correlates of talker intelligibility for adults and children. J. Am. Acad. Audiol. 116 (5), 3108–3118.
Hazan, V., Simpson, A., 1996. Cue-enhancement strategies for natural VCV and sentence materials presented in noise. Speech Hear. Lang. 9, 43–55.
Junqua, J., 1993. The Lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am. 93, 510–524.
Koutsogiannaki, M., Pettinato, M., Mayo, C., Kandia, V., Stylianou, Y., 2012. Can modified casual speech reach the intelligibility of clear speech? In: Interspeech 2012, Portland, Oregon, USA.
Krause, J., Braida, L., 2004a. Acoustic properties of naturally produced clear speech at normal speaking rates. J. Acoust. Soc. Am., 115.
Krause, J., Braida, L., 2004b. Investigating alternative forms of clear speech: the effects of speaking rate and speaking mode on intelligibility. J. Acoust. Soc. Am. 112 (5), 2165–2172.
Langner, B., Black, A., 2005. Improving the understandability of speech synthesis by modeling speech in noise. In: ICASSP, pp. 265–268.
Lecumberri, M.C.M.G., 2012. The intelligibility of Lombard speech for non-native listeners. J. Acoust. Soc. Am. 132, 1120–1129 (Letters to the Editor).
Leino, T., Laukkanen, A., Radolf, V., 2011. Formation of the actor's/speaker's formant: a study applying spectrum analysis and computer modeling. J. Voice.
Liu, S., Zeng, F., 2006. Temporal properties in clear speech perception. J. Acoust. Soc. Am. 120 (1), 424–432.
Lombard, E., 1911. Le signe de l'elevation de la voix, annals maladiers oreille. Larynx Nez Pharynx 37, 101–119.
Lu, Y., 2009. Production and Perceptual Analysis of Lombard Effect. Department of Computer Science, The University of Sheffield (Ph.D. thesis).
Lu, Y., Cooke, M., 2008. Speech production modifications produced by competing talkers, babble and stationary noise. J. Acoust. Soc. Am., 3261–3275.
Lu, Y., Cooke, M., 2009. The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. Speech Commun., 1253–1262.
McLoughlin, I., Chance, R.J., 1997. LSP-based speech modifications for intelligibility enhancement. In: 13th Int. Conf. DSP, pp. 591–594.
Mohammadi, S., Kain, A., van Santen, J., 2012. Making conversational vowels more clear. In: Interspeech, Portland, Oregon, USA.
Niederjohn, R., Grotelueschen, J., 1976. The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. IEEE Trans. Audio Speech Lang. Process. 24, 277–282.
Picheny, M.A., Durlach, N.I., 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., 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.
Quatieri, T., McAulay, R., 1991. Peak-to-rms reduction of speech based on a sinusoidal model. IEEE Trans. Signal Process. 39, 273–288.
Raitio, T., Suni, A., Vainio, M., Alku, P., 2011. Analysis of HMM-based Lombard speech synthesis. In: Interspeech, pp. 2781–2784.
Rhebergen, K.S., Versfeld, N.J., Dreschler, W.A., 2006. Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. J. Acoust. Soc. Am. 120, 3988–3997.
Roebel, A., Rodet, X., 2005. Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In: Digital Audio Effects (DAFx)., pp. 30–35.
Sauert, B., Vary, P., 2006. Near end listening enhancement: speech intelligibility improvement in noisy environments. In: ICASSP, pp. 493–496.
Skowronski, M.D., Harris, J.G., 2006. Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. Speech Commun. 48, 549–558.
Summers, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., Stokes, M.A., 1988. Effects of noise on speech production: acoustical and perceptual analyses. J. Acoust. Soc. Am. 84, 917–928.
Tang, Y., Cooke, M., 2011. Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In: Interspeech, pp. 345–348.
Valbret, H., Moulines, E., Tubach, J., 1992. Voice transformation using PSOLA technique. Speech Commun. 11, 175–187.
Valentini-Botinhao, C., Yamagishi, J., King, S., Florence, Italy, 2011. Can objective measures predict the intelligibility of modified hmm-based synthetic speech in noise? In: Interspeech 2011.
Womack, B., Hansen, J., 1996. Classification of speech under stress using target driven features. Speech Commun. 20, 131–150.
Zorila, T., Kandia, V., Stylianou, Y., 2012. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In: Interspeech 2012, Portland, Oregon, USA.