
Article:
Marxer, R., Barker, J.P. orcid.org/0000-0002-1684-5660, Alghamdi, N. et al. (1 more author) (2018) The impact of the Lombard effect on audio and visual speech recognition systems. Speech Communication, 100. pp. 58-68. ISSN 0167-6393

https://doi.org/10.1016/j.specom.2018.04.006

Speech Communication 100 (2018) 58–68

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

The impact of the Lombard effect on audio and visual speech recognition systems

Ricard Marxer*[,a,b], Jon Barker[b], Najwa Alghamdi[b], Steve Maddock[b]

[a] Université de Toulon, Aix Marseille Univ, CNRS, LIS, Marseille, France
[b] Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK

ARTICLE INFO

Keywords:
Lombard speech
Multimodal speech
Automatic speech recognition
Intelligibility
Robust speech processing
Visual speech

ABSTRACT

When producing speech in noisy backgrounds talkers reflexively adapt their speaking style in ways that increase speech-in-noise intelligibility. This adaptation, known as the Lombard effect, is likely to have an adverse effect on the performance of automatic speech recognition systems that have not been designed to anticipate it. However, previous studies of this impact have used very small amounts of data and recognition systems that lack modern adaptation strategies. This paper aims to rectify this by using a new audio-visual Lombard corpus containing speech from 54 different speakers – significantly larger than any previously available – and modern state-of-the-art speech recognition techniques.

The paper is organised as three speech-in-noise recognition studies. The first examines the case in which a system is presented with Lombard speech having been exclusively trained on normal speech. It was found that the Lombard mismatch caused a significant decrease in performance even if the level of the Lombard speech was normalised to match the level of normal speech. However, the size of the mismatch was highly speaker-dependent thus explaining conflicting results presented in previous smaller studies. The second study compares systems trained in matched conditions (i.e., training and testing with the same speaking style). Here the Lombard speech affords a large increase in recognition performance. Part of this is due to the greater energy leading to a reduction in noise masking, but performance improvements persist even after the effect of signal-to-noise level difference is compensated. An analysis across speakers shows that the Lombard speech energy is spectro-temporally distributed in a way that reduces energetic masking, and this reduction in masking is associated with an increase in recognition performance. The final study repeats the first two using a recognition system training on visual speech. In the visual domain, performance differences are not confounded by differences in noise masking. It was found that in matched-conditions Lombard speech supports better recognition performance than normal speech. The benefit was consistently present across all speakers but to a varying degree. Surprisingly, the Lombard benefit was observed to a small degree even when training on mismatched non-Lombard visual speech, i.e., the increased clarity of the Lombard speech outweighed the impact of the mismatch.

The paper presents two generally applicable conclusions: i) systems that are designed to operate in noise will benefit from being trained on well-matched Lombard speech data, ii) the results of speech recognition evaluations that employ artificial speech and noise mixing need to be treated with caution: they are overly-optimistic to the extent that they ignore a significant source of mismatch but at the same time overly-pessimistic in that they do not anticipate the potential increased intelligibility of the Lombard speaking style.

1. Introduction

Automatic speech recognition is now finding widespread application in everyday environments. For example, distant microphone systems designed for household use are becoming increasingly common (Google Home, Amazon Alexa, etc.). In typical everyday scenarios there can be high levels of interfering background noise present. This causes significant challenge for recognition systems. Even if the interfering noise can be well anticipated, the performance of a speech recogniser (whether machine or human) can be degraded by the simple fact that the noise will energetically mask part of the speech signal. This noise masking leads to a loss of phonetic information and hence an increase in recognition errors. Fortunately, speakers are sensitive to this effect and, in challenging communication settings, they reflexively adapt their speech production in ways that counter the effects of noise masking. This adaptation which includes an increase in signal energy, a tilt of the

* Corresponding author.
E-mail addresses: ricard.marxer@lis-lab.fr, marxer@univ-tln.fr (R. Marxer).

https://doi.org/10.1016/j.specom.2018.04.006
Received 14 July 2017; Received in revised form 12 April 2018; Accepted 18 April 2018
Available online 20 April 2018

speech spectrum and an increase in vowel duration, has become known as the Lombard effect, named after Étienne Lombard who first described it in 1909 (Lombard, 1911; Brumm and Zollinger, 2011). This effect will be present to greater or lesser extent whether humans are conversing with a human partner or with an automatic recognition system. However, whereas the human listener is naturally able to exploit the Lombard speaking style to better understand speech in noise, an automatic speech recogniser may gain no benefit. Indeed, depending on how the system has been designed and trained, Lombard speech may even have a negative impact on performance.

Whereas there have been many studies of the Lombard effect from the perspective of human-human communication, there has been surprisingly little examination of its consequences on automatic speech recognition performance (Huang and Chen, 2001). In fact, many of the formative studies of noise-robust speech recognition have chosen to totally disregard the effect by employing speech recorded in studio conditions to which noise has been artificially added after recording, e.g., Aurora 2 (Hirsch and Pearce, 2000), Aurora 4 (Parihar et al., 2004), CHiME-1 (Barker et al., 2013) and CHiME-2 (Vincent et al., 2013). The few exceptions (e.g., Junqua, 1993; Hansen and Varadarajan, 2009) have used very small collections of Lombard speech and have been conducted without the benefit powerful speaker adaptation techniques (such as speaker adaptive training (Anastasakos et al., 1997)) that are now part of the standard automatic speech recognition pipeline.

The purpose of this paper is to examine the impact of the Lombard effect in isolation from the other difficulties of noise-robust speech recognition. The work employs a multimodal corpus in which headphone noise presentation has been employed to collect Lombard and non-Lombard speech from 54 individual speakers. The paper is arranged as three separate studies. The first study considers Lombard speech as a source of mismatch. If a system has been trained using regular speech artificially mixed with noise, how well will it be expected to perform when encountering Lombard speech? The second study examines the potential for the Lombard effect to improve recognition performance. If Lombard speech is well-modelled does it allow for better speech recognition performance in noise? The final study examines the Lombard effect from a visual speech perspective. Are the effects of the mismatch equally present in the visual domain? Does the more pronounced articulation of Lombard speech allow for better visual speech recognition in matched conditions?

The remainder of this paper is structured as follows. Section 2 reviews previous studies of Lombard speech. We summarise the main characteristics of Lombard speech and review the impact of the Lombard effect on speech intelligibility. Section 3 presents the Lombard speech materials that are used for the studies in the paper. Sections 4–6 present the mismatched, matched and visual Lombard studies, respectively. The paper concludes with a summary of major findings in Section 7.

2. Background

When speaking in the presence of background noise, talkers will increase their vocal effort so that their speech remains intelligible. The changes that are observed are complex and are likely due to the effect of multiple mechanisms. The classic 'Lombard effect' (Lombard, 1911) is typically described as an involuntary response and is believed to be primarily mediated via self-monitoring of the voice, i.e., if the talker is unable to hear their own voice then their vocal effort will reflexively increase (Svirsky et al., 1992). In addition to self-monitoring, talkers will also naturally adapt their speech behaviour if they detect that they are not intelligible to their conversational partner. In situations where the receiver is struggling to understand, a talker will adopt a so-called 'clear speech' style regardless of whether there is noise present in the environment (Picheny et al., 1986), for example, when speaking to non-natives or hearing-impaired listeners. When conversing in noise, both classic Lombard adaptations and clear speech adaptations are likely to be co-occurring. There is debate in the literature as to whether self-monitoring or perceived intelligibility is the primary factor driving speech adaptations. In this paper we follow the classic study of Junqua (1993) in emphasising the former by using masking noise to induce the effect while talkers read sentence lists.

There have been many previous studies that have attempted to characterise the Lombard effect (Summers et al., 1988; Stanton et al., 1988; Hansen, 1989; Junqua, 1993). Although the findings of these studies have differed in detail, a consistent description of Lombard speech has emerged: Spectral effects include an increase in fundamental frequency, a tilting of the spectrum that emphasises higher frequencies and a shift in formant center frequencies (particularly an increase of F1). There is also a pronounced increase in overall sound energy: Junqua (1993) reports an increase of 15 dB for speech produced in 85 dB white noise; Pittman and Wiley (2001) report a 14.5 dB increase with 80 dB inducing noise. In the temporal domain the main effect is an increase in vowel duration leading to an overall reduction in speech rate. This effect has been observed to have a linguistic dependency: the vowel lengthening is greater in content words than in function words (Patel and Schell, 2008). Visually it has been observed that Lombard speech exhibits both larger facial movements (e.g., lip and jaw) and more pronounced rigid-head motions (Summers et al., 1988).

Despite the complexity of Lombard adaptations, there is no evidence that the effect is actively adapted to the characteristics of the inducing noise, i.e., no systematic effects have been observed in studies that have directly compared inducing noises with different levels or spectral properties (Lu and Cooke, 2009; Garnier and Henrich, 2014). This is perhaps logical considering that the properties of the noise at the ear of the receiver are likely to be different to those at the ear of the talker. Further, all studies have also found large variabilities between speakers, with additional differences between genders (Junqua, 1993). These variabilities mean that it is not possible to accurately anticipate a person's Lombard speech from knowledge of their plain speech.

The Lombard reflex is highly effective in increasing the intelligibility of speech in noise. This is perhaps unsurprising given the large increases in speech energy which will boost the effective SNR. However, a significant intelligibility gain remains even after the intensity difference between the plain speech and Lombard speech styles is removed (Summers et al., 1988; Pittman and Wiley, 2001). Lu and Cooke (2008, 2009) have shown that these intelligibility increases are well predicted by models of energetic masking. In particular, in Lombard speech masking is reduced by the change in spectral tilt and the increased vowel duration. When spectral and durational effects are decomposed by mapping just one or the other onto normal speech, it is found that the intelligibility benefit can be predominantly attributed to the spectral differences (Cooke et al., 2014).

There have been fewer studies of the intelligibility benefits of *visual* Lombard speech. Vatikiotis-Bateson et al. (2006) found that, when presented in noise, visual information improved the intelligibility for both plain speech and Lombard speech. However, they observed no additional benefit for the visual Lombard condition. This is in contradiction to more recent studies (Kim et al., 2011; Fitzpatrick et al., 2015) where the Lombard visual benefit has been seen to be larger than the plain speech visual benefit. The authors proposed that this was partly due to an increase in phonetic information provided by the visual speech signal. In support of this, Fitzpatrick et al. (2013) examined Lombard speech lip-reading in a visual-only condition and noted a significant Lombard benefit for lip-reading both vowels and consonants.

The fact that Lombard speech is more intelligible would suggest that if modelled correctly it should also afford higher performance in automatic speech recognition systems. However, as discussed by Junqua (1993), if a recognition system is trained on plain speech then an unanticipated Lombard mismatch will cause severe degradation in performance. Various Lombard compensation techniques have been proposed and are reviewed by Hansen and Varadarajan (2009). These

include traditional cepstral means normalization (CMN) (Bou-Ghazale and Hansen, 2000); spectral-slope dependent weighting metrics (Stanton et al., 1989); linear transformation in the LPC cepstral domain (Wakao et al., 1996); retraining with synthesised Lombard speech (Hansen and Bou-Ghazale, 1995). Hansen has developed a source-filter framework for stressed speech modelling which has been used to develop a number of compensation methods (Hansen and Clements, 1995).

In our paper, we are less concerned with adaptation techniques per se but rather we are concerned with the performance of recognition systems in non-adapted and well-adapted conditions. In Study I we will consider a Lombard utterance being processed by a system that has been trained using plain speech of the same speaker. In Study II, we assume access to matched Lombard training data and use modern training and adaptation techniques to make high quality matched Lombard speech models. We are interested to observe the impact of Lombard speech on recognition performance once the mismatch has been corrected and how this varies over speaker and presentation SNR. We are also interested in the interaction between energetic unmasking and changes to intrinsic intelligibility due to increased or decreased phonetic discriminability. This point is reinforced by Study III which uses visual-only Lombard speech for which recognition performance is directly related to intrinsic intelligibility (there is no masking).

3. Materials: the Lombard Grid corpus

Speech materials have been taken from the Audio-Visual Lombard Grid corpus[1] (Alghamdi et al., 2018). Essential details for the data collection are repeated here for the sake of completeness.

Speech materials were collected from 54 paid adult volunteers (24 male, 30 female). All volunteers were native English speakers with no self-reported hearing problems.

Each stimulus consisted of a six-word sentence such as 'place blue at A 2 please' following the Grid corpus sentence syntax (Cooke et al., 2006): < command: 4 > < colour: 4 > < preposition: 4 > < letter: 25 > < digit: 10 > < adverb: 4 > , where the number of choices for each keyword is indicated in the angle brackets (and there are 25 letters since the multisyllabic letter 'W is not included). There are 64,000 possible unique sentences with this structure. 34,000 of these have been used in the Grid corpus. Sentences for the Lombard Grid corpus are drawn at random from the 30,000 remaining sentences unused by Grid.

Speech was collected in both a Lombard (L) and non-Lombard (NL) condition. All recordings were conducted in an Industrial Acoustics Company (IAC) single-walled acoustically-isolated booth. Participants were prompted with the sentences to read using a simple interface. The Lombard effect was induced by headphone presentation of speech-shaped noise at a level of 80 dB SPL. The speech-shaped noise was constructed by filtering white noise to match the long term spectrum of an adult male talker.

For each participant a different set of 50 unique utterances was recorded.[2] The same set was recorded in both the L and NL conditions. For each condition the 50 utterances were separately randomised and then arranged into 5 blocks of 10 utterances. The L and NL blocks were presented in an alternating pattern. Recordings were completed over two 5-block sessions. Each block of 10 utterances was preceded by 5 'warm up' utterances in the same condition. These utterances were discarded after recording. The Lombard-inducing noise was played continuously throughout the recording of the Lombard blocks with the noise being automatically turned on and off between blocks under software control.

The closed-design of the headphones causes a degree of own-voice attenuation. In the study of Lu and Cooke (2008) this small attenuation was observed to have no significant effect on a set of basic acoustic measurements. However, as a precautionary step we compensated for the attenuation by playing the speaker's own voice back to them through the headphones during recording. The level of playback was carefully adjusted so that the experience of talking with and without headphones was perceptually identical.

The participant's utterances were monitored by the experimenter during recording. The interface which presented the prompt could also allow the experimenter to ask the participant to repeat any sentences that were misspoken. This had the double benefit of ensuring the quality of the data, but also of putting the participant in a natural communication setting.

The speech signals were recorded with an AKG C414-XLS condenser microphone placed 30 cm in front of the talker and digitized via a MOTU 8-pre 16 × 12 Audio Interface. The audio is distributed as 16 bit, 16 kHz wave files. Both frontal and profile video were recorded. The video was captured using head-mounted cameras so that the speaker's head pose remains fixed throughout recording hence allowing precise comparison of the L and NL conditions (full details of the headset design are presented in Alghamdi et al. (2018). The audio and visual channels for each recording session were precisely aligned. The audio channel was then used to end-point the utterances. Utterances were then segmented from the audio and visual channels with a 200 ms margin around the endpoints. This margin was included to accommodate anticipatory visual speech cues.

In summary, the speech materials consist of 5400 segmented full-face video, profile video and audio signals each representing a single sentence. There are 2700 unique utterances spoken in a Lombard condition and 2700 corresponding non-Lombard reference utterances (i.e., the same sentence spoken by the same speaker).

4. Study I: the Lombard effect as a source of mismatch

4.1. Motivation

In human-human communication the Lombard effect is generally a helpful adaptation that protects speech against the deleterious effects of noise. It reduces masking and exaggerates differences between acoustically confusable phonemes. However, in an automatic speech recognition context it can also have a negative impact. If the system has not been trained on Lombard speech, then the mismatch between the test data and statistical models of non-Lombard speech can lead to an increase in recognition errors.

The purpose of this first study is to measure the impact of the effect of the mismatch and to investigate how it varies across speakers. The study assumes a 'worst case' scenario in which a system has been trained by mixing studio recorded speech with artificially added noise, i.e., the style of training used in early robust ASR evaluations such as Aurora 2 (Hirsch and Pearce, 2000) and Aurora 4 (Parihar et al., 2004). The system then encounters a Lombard utterance in a matching noise background. It is assumed that there is no Lombard adaptation data, so that the only normalisations or adaptations allowed must operate using just the utterance that is to be transcribed.

4.2. Methods

Training and testing data have been constructed by artificially adding speech shaped noise to the Lombard corpus utterances. Noise levels have been chosen so that the non-Lombard utterances have SNRs ranging from −12 dB to 0 dB in steps of 3 dB. This range was chosen to ensure coverage of noise levels that were free from performance floor and ceiling effects – it also spans the −9 dB SNR used in the Lombard Grid sentence perception studies of Lu and Cooke (2008). Note that because the Lombard utterances have greater energy, for a given noise

[1] http://spandh.dcs.shef.ac.uk/lombardgrid/.

[2] Except for two pairs of speakers, #6-#29 and #25-#26, where each pair read the same sentence list.

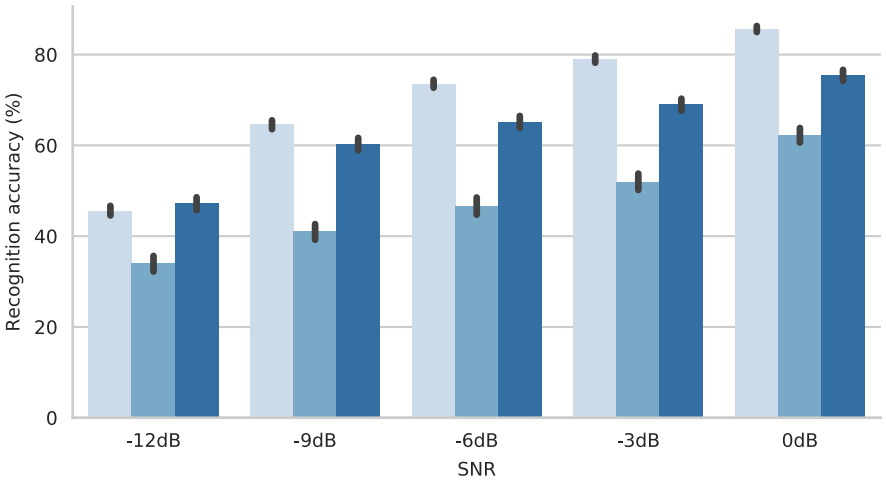

Fig. 1. Recognition accuracies for different SNR values and testing speech conditions. All models have been trained on non-Lombard speech. From light to dark: non-Lombard, Lombard, 'compensated' Lombard. The height indicates the average across utterances while the black line depicts the standard error (95% confidence interval).

level, they will have a significantly higher SNR than the non-Lombard mixtures. This energy difference is the biggest part of the mismatch. In order to measure how much of the mismatch effect is due to this alone, a further set of 'compensated' Lombard (CL) noisy utterances are generated in which the Lombard utterances are normalised to the same normalisation energy as the non-Lombard utterances before adding the noise, i.e., this set of noisy Lombard utterances will be at an SNR that matches the non-Lombard data.

The recording set-up uses a desktop microphone which does not allow the speaker-to-microphone distance to be carefully controlled. This introduces an artificial level variability in the signals. In the original Grid corpus recordings, this variability was removed by returning all signals to the same root mean square (RMS) level. A similar strategy is employed here, but with care taken to make sure that the non-Lombard versus Lombard level difference is not also removed. In detail, the non-Lombard and 'compensated' Lombard signals are all normalised to a fixed root mean square (RMS) amplitude value of 0.05. All the Lombard utterances of a particular recording session of each speaker are then scaled to an RMS value of $0.05 \cdot \overline{x}^{\mathrm{L}}_{\mathrm{rms}} / \overline{x}^{\mathrm{NL}}_{\mathrm{rms}}$, where $\overline{x}^{\mathrm{L}}_{\mathrm{rms}}$ and $\overline{x}^{\mathrm{NL}}_{\mathrm{rms}}$ are the average RMS values of the Lombard and non-Lombard speech utterance of the session respectively. This way all utterances from the same session will be scaled equally and the per-session average RMS ratios between the Lombard and non-Lombard speech will remain the same as in the original recording.

Speech recognition experiments have been performed using an auditory spectrogram representation. Specifically, signals are filtered with a bank of 32 overlapping Gammatone filters, with centre frequencies uniformly spaced on the equivalent rectangular bandwidth (ERB) scale between 100 Hz and 7 kHz (Glasberg and Moore, 1990). The instantaneous envelope of each Gammatone filter is computed using a Hilbert transform. The envelope is then sampled at a 10 ms frame shift and the sampled values are log-compressed to obtain an approximation to the auditory nerve firing rate - the 'ratemap spectro-temporal representation (Brown and Cooke, 1994). Similar filterbank representations have been shown to afford state-of-the-art recognition results on the Grid corpus in previous studies (Meutzner et al., 2017) and the particular representation described here has been used to form the basis of successful models of speech intelligibility in noise (Cooke, 2006).

In all studies ASR systems have been trained using the Kaldi toolkit (Povey et al., 2011). The ASR is performed with a typical GMM-HMM setup. Each feature frame is concatenated with 3 neighbours from either side. The dimensionality is then reduced to 40 by means of a Linear Discriminative Analysis (LDA) transform. The resulting feature vectors are mapped to clusters of tied context-dependent sub-phonetic triphone states using Gaussian mixture models. Following Young et al. (1994) the states are tied using decision trees with a maximum of 2500 clusters and 15,000 Gaussian components. During decoding the language model assigns a uniform probability distribution among all possible Grid sentences, while at training time the language model assigns all the probability to the uttered Grid sentence, effectively performing a forced alignment.

The training procedure follows the widely-used Kaldi GMM-HMM recipe consisting of a sequence of alternating model estimation and alignment steps. The stages, of increasing model complexity, are: i) a monophone model using the raw features and a uniform alignment, ii) clustered triphone states model using features and their time derivatives, iii) splicing of features and LDA transform with maximum likelihood linear transform (MLLT) model adaptation, and iv) speaker adaptive training (SAT) with feature-space maximum likelihood linear regression (fMLLR). During testing, the fMLLR transform is estimated for each utterance independently, thus not using any speaker information.

SNR-specific models have been used throughout the experiments. This setup allows setting the focus on the consequences of speech style without introducing errors due to SNR variability. In order to remove the effects of dataset mismatch and maintain a balance between speaking styles only speech data from the aforementioned Lombard Grid Corpus was used for training and testing the ASR systems. A multi-speaker setup was chosen over training speaker-dependent models. The use of SAT and fMLLR allowed us to maximise the data available for training whilst increasing the specificity of the models to each speaker. Furthermore the scenario is common in many existing speech recognition tasks and applications.

Due to the relatively small amounts of speech data, a leave-one-out style of training and testing is employed: 50 separate cuts of the data are employed, in which one Lombard/non-Lombard utterance pair from each of the 54 speakers is used to define a 54 utterance Lombard and non-Lombard test set which is evaluated against a model trained on the remaining 54×49 non-Lombard utterances. This selection is then rotated around until all 50 utterances for each speaker have been evaluated in both conditions. Following previous Grid corpus ASR studies (e.g., Barker et al., 2013), results are then presented as percentage of keywords (letter and digit) recognised correctly, either averaged across all speakers for each condition, or analysed on a per-speaker basis.

4.3. Results

Fig. 1 shows the average[3] keyword correctness across all speakers for models trained on non-Lombard speech and tested on each of the conditions: non-Lombard, Lombard and 'compensated' Lombard. The

[3] Error bars in this and all other plots represent a 95% confidence interval.

Table 1
Keyword (digit & letter) recognition accuracy for models trained on non-Lombard speech data and tested on different speech conditions.

	−12dB	−9dB	−6dB	−3dB	0dB
Non-Lombard	45.6%	64.5%	73.6%	79.0%	85.6%
Lombard	35.2%	42.9%	48.8%	54.4%	64.6%
'compensated' Lombard	47.2%	60.2%	65.2%	69.1%	75.7%

same data is shown in Table 1. In the matched non-Lombard condition performance falls steadily from over 80% at 0 dB to 46% at −12 dB. The Lombard mismatch results in a decrease in correctness in the range of 20–30%, a degradation in performance similar to that observed when decreasing SNR by 9 dB. In the 'compensated' condition, in which the training data and test data have identical SNRs, there remains a degradation in performance, albeit of a smaller magnitude at around 5–10%. Note, the degradation disappears at the lowest SNR of −12 dB. It is likely that in this extreme condition, the remaining differences between the normal and 'compensated' Lombard are largely concealed by the extensive energetic masking.

It may seem counter-intuitive that the 'compensated' Lombard utterances (which have a lower SNR) are better recognised than the less heavily masked Lombard utterances. To understand this it is helpful to consider the *SNR mismatch* as a separate component of the total mismatch. For the Lombard utterances the SNR is poorly matched to that of the non-Lombard mixtures so performance will be poor despite the fact that the SNR is higher. Further, although they are less heavily masked, the extra information available has no value because it was not observed in the more heavily masked training data. In the compensated case, the SNR is reduced but so is the SNR mismatch, so overall performance improves.

The average results conceal a large amount of speaker variability. Fig. 2 shows a Box-plot of per-speaker percentage changes between non-Lombard and Lombard recognition scores (light) and between non-Lombard and 'compensated' Lombard recognition scores (dark). Negative scores show a degradation due to mismatch. In the non-compensated case, the Lombard mismatch degrades performance for all speakers at all SNRs – albeit by widely varying amounts. However, for some speakers this degradation is fully recovered by the simple gain compensation (i.e., the dark bars extend up to 0% change).

4.4. Discussion

It has been seen that without compensation the Lombard effect can cause very significant mismatches. This is in agreement with the findings of Junqua (1993) and Hansen (1989). The impact on performance is likely to be task and data-set dependent. Some features of Grid will make it sensitive to mismatch. In particular, the regularity of the sentences and the read-speech speaking style leads to acoustic models with narrow variance. This is compounded by the level normalisation that was performed to remove channel variability, but which will also have removed natural speaker level variability. However, the within-speaker level variability of normal read speech is very small compared with the large differences between normal and Lombard speech (6.5 dB on average). In general though, the Grid task with a small and readily-discriminable vocabulary is intrinsically robust. The impact of Lombard mismatch is likely to be more severe when measuring error rates in tasks with larger vocabularies.

The most notable feature of the results is the very large variability of the mismatch when viewed across speakers. This is consistent with previous studies which have shown that the degree of the Lombard effect is highly speaker and situation dependent. In the current study this is readily observed when looking at the energy gain between normal and Lombard speech. Measuring across all utterances and all speakers the average gain was 6.5 dB. However, examining the distribution of gains across speakers (Fig. 3) it is seen that it varies from 3 dB to as large as 13 dB. This is remarkable considering that for each talker these gains represent an average computed from 50 sentences of each condition, i.e., the variability cannot be explained by the effects of outlying utterances. This large speaker difference sheds some light on the large deviation in averages across speakers reported in the literature summarised in Table 2. The apparent inconsistency of these studies is perhaps not surprising given some have used as few as two speakers and that they have typically used much smaller amounts of speech material than used in the current work.

5. Study II: the Lombard effect and intelligibility

5.1. Motivation

Previous human listening studies have consistently shown that the Lombard effect significantly increases the intelligibility of speech in noise (Summers et al., 1988; Junqua, 1993). It would therefore seem likely that in ASR systems, Lombard speech can be recognised with lower error rates than normal speech if the systems are correctly adapted to avoid mismatch problems. This is likely to be true for the trivial reason that the Lombard effect introduces a gain that increases SNR and reduces energetic masking. However, benefits may remain even after the Lombard gain has been removed, i.e., it is possible that Lombard speech affords better recognition results than normal speech

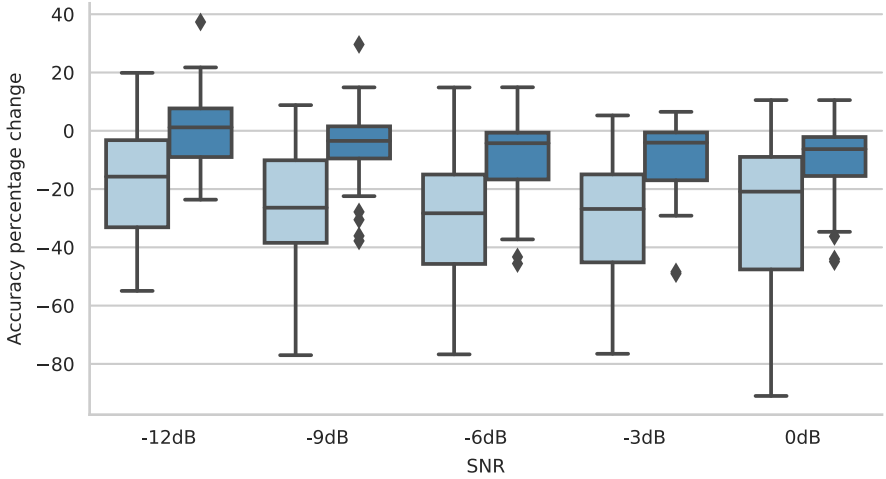

Fig. 2. Per-speaker accuracy percentage change between non-Lombard and Lombard (light) or 'compensated' Lombard (dark) for different SNR values. All models have been trained on non-Lombard speech.

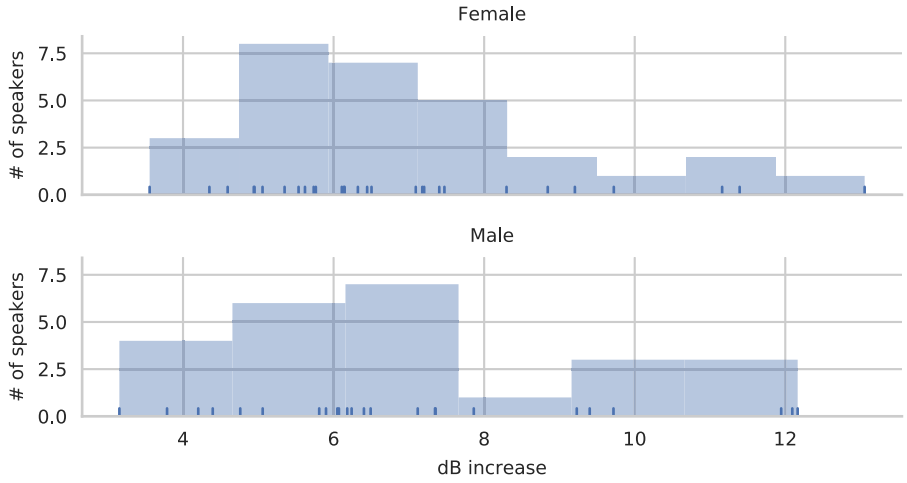

Fig. 3. Distribution of per-speaker average energy gains in dB between non-Lombard and Lombard speech. Each dark tick on the x-axis represents the average energy gain of an individual speaker.

Table 2
Summary of level increases observed in previous studies.

Study	Level	Noise type	Subjects	Energy gain
Summers et al. (1988)	80 dB	white noise	2 male	4.6 dB
Junqua (1993)	85 dB	white noise	5 m / 5 f	15 dB
Tartter et al. (1993)	80 dB	white noise	2 female	3.7 dB
Pittman and Wiley (2001)	80 dB	WBN, babble	5 female	14.5 dB
Lu and Cooke (2008)	82 – 96 dB	SSN, babble	4 m / 4 f	3 – 9 dB

even when the SNRs of the two styles are held constant. This has been observed in listening studies and has been partly attributed to the spectral modifications reducing energetic masking (Summers et al., 1988; Junqua, 1993). Lombard speech may also be more intelligible to the extent that it shares hyper-articulation characteristics of clear speech that will tend to enhance phonetic contrasts and improve discriminability.

This study considers a 'best case' processing of Lombard speech by training ASR systems in matched Lombard conditions. The study first compares non-Lombard and Lombard systems trained with equal amounts of additive noise in which the Lombard signal will have a higher SNR. We then compare the performance of these systems to that of systems trained with Lombard speech mixed at the same SNR as the normal speech. Given the large speaker variabilities observed in Study I, we are particularly interested in the spread of effect across speakers. If Lombard speech is easier to recognise independently of the effect of SNR, then is this consistently observed across speakers? This question is significant because it is possible that noise robustness studies that have previously used artificial mixing (i.e., with no account for the Lombard effect) may have been *overestimating* the difficulty of real speech recognition.

In this study we also attempt to relate the variability in recognition performance to a model of energetic masking that has been shown to be a good predictor of human speech intelligibility in noise (Cooke, 2006). In particular, when comparing Lombard speech recognition performance to normal speech recognition performance can we factor out the impact of masking from the impact of changes to the acoustic model?

5.2. Methods

The experimental setup used in this study is similar to that employed in Study I. Namely, for each noise level, 50 splits of data are created for each of the three speech conditions: non-Lombard, Lombard and 'compensated' Lombard. For each split a model is trained on 54(*speakers*) × 49 utterances and is tested on the left-out 54 utterances (one from each speaker) at the same noise level. Results are then pooled across all splits. The model structure and training regime are identical to that used in Study I. The difference is that separate sets of models are trained for each of the three major conditions and the test data is recognised using models matched for noise-level, speaker and condition.

5.3. Results

Fig. 4 displays the keyword accuracy for non-Lombard (light), Lombard (medium) and 'compensated' Lombard (dark) for each of the SNRs. The results are repeated in Table 3. Note that the results for the matched non-Lombard condition repeat those from the mismatch Study I in which all systems were trained with non-Lombard data.

It can be observed that for models trained in matched conditions, Lombard speech obtains an accuracy equivalent to that of non-Lombard speech at around 6 dB higher SNR. A large part of this increase will be simply due to the SNR improvement due to the Lombard gain which was previously measured as 6.5 dB on average. However, even after compensating for the energy differences between Lombard and non-Lombard speech, there remains an improvement in performance at the lowest SNR equivalent to about 3 dB of noise reduction.

5.4. Discussion

The increased recognition performance observed for the 'compensated' Lombard condition shows that the improvement in recognition benefit of the Lombard speaking style cannot be solely attributed to the increased energy, i.e., performance improves even when the signals are normalised to the original speech level prior to addition of noise. This is likely to be due to unmasking caused by redistribution of the energy in the spectrum. To examine this effect we have used Cooke's Glimpse Proportion (GP) measure of energetic masking (Cooke, 2006) which has previously been shown to be a good predictor of speech intelligibility in human listening experiments. The GP is defined as the proportion of time-frequency cells in the auditory spectrogram representation in which the energy of speech is locally 3 dB greater than the energy of noise. This can be readily computed using the pre-mixed speech and noise signals.

Fig. 5 plots the average GP across all speakers for utterances at different SNRs and for each of the Lombard, 'compensated' and normal speech conditions. These averages closely follow the pattern of the matched-training recognition performances shown in Fig. 4. The GP for Lombard speech is increased by an amount consistent with the 6.5 dB average gain. Further, the GP of the level 'compensated' Lombard speech is also increased, by an amount equivalent to a 3 dB noise

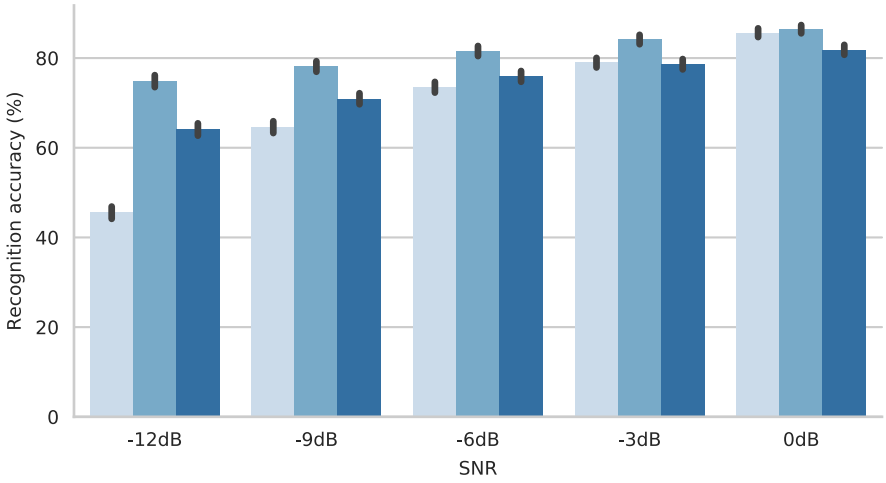

Fig. 4. Recognition accuracies for different SNR values and testing speech conditions. All models have been trained on matched conditions. From light to dark: non-Lombard, Lombard, 'compensated' Lombard. The height indicates the average across utterances while the black line depicts the standard error.

Table 3
Keyword (digit & letter) recognition accuracy tested on different speech conditions for models trained on matched speech data.

	−12dB	−9dB	−6dB	−3dB	0dB
Non-Lombard	45.6%	64.5%	73.6%	79.0%	85.6%
Lombard	75.9%	80.2%	83.7%	85.6%	88.3%
'compensated' Lombard	64.1%	71.0%	76.1%	78.9%	82.0%

reduction.

Fig. 6 examines the relation between GP and recognition performance in greater detail. Here, for each of the three data sets, we show a scatter plot of GP versus recognition accuracy with each point representing a speaker in one SNR condition. In the non-Lombard speech the relation between GP and correctness is clearly seen and this mirrors similar data seen in previous analyses of the larger Grid corpus (Barker and Cooke, 2007). There is a strong relation between performance and glimpse proportion. As the GP increases from near 0% the accuracy increases rapidly. When considering Lombard speech and the 'compensated' Lombard speech, it can be seen that although the shift of points towards higher GP is accompanied by a commensurate shift towards higher recognition accuracies, in detail, the picture is more complicated. The spread of performances across speakers is much wider with some speakers achieving greater recognition performance benefits than the decrease in masking would predict, and others achieving little benefit. This suggests that there is a secondary effect which we might name 'intrinsic intelligibility' that for some speakers has more influence than pure masking alone.

In order to make the effects of transitioning from the normal to the Lombard style more explicit, Fig. 7 shows the movements of individual speakers in the GP versus accuracy space. The upper row shows results for uncompensated Lombard where large improvements in accuracy are associated with big increases in GP. The lower row shows the same plot for 'compensated' Lombard. What becomes apparent is that the relation between masking release and performance improvement is strongest at the lowest SNRs where the non-Lombard utterances start with very low GPs. At the intermediate SNR of −6 dB the masking release is still present but the performance improvements are not consistently observed after gain compensation. Note, although it appears that performance actually decreases for some speakers, each of these transitions is based on just 100 tokens and so individual changes have to be 10–15% before becoming statistically significant. However, at the highest SNR, there is a large cluster of speakers who start with a low recognition accuracy and then show a decline in recognition performance in the 'compensated' Lombard condition. The statistically significant drop in performance observed in Fig. 4 appears to be concentrated in these speakers. It remains unclear whether this is a genuine reduction in the 'intrinsic intelligibility' of these speakers or whether it is a consequence

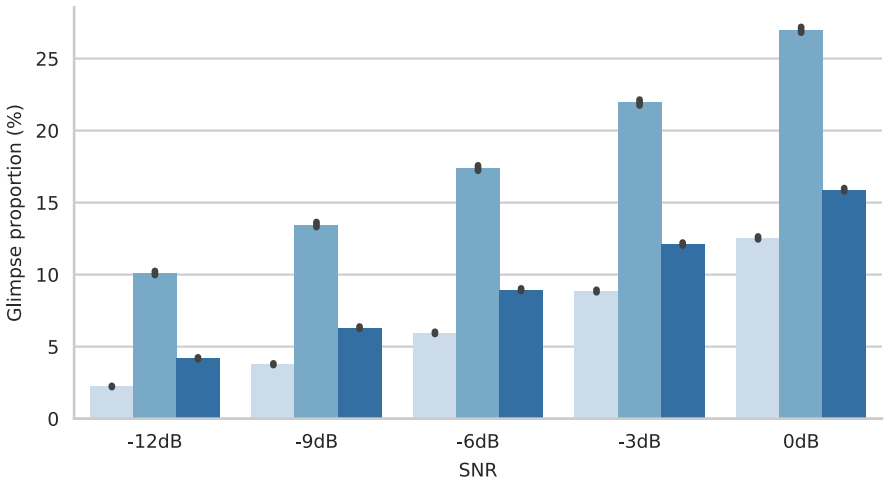

Fig. 5. Average Glimpse Proportion (GP) of the utterances at different SNRs for the different speech conditions. From light to dark: non-Lombard, Lombard, 'compensated' Lombard.

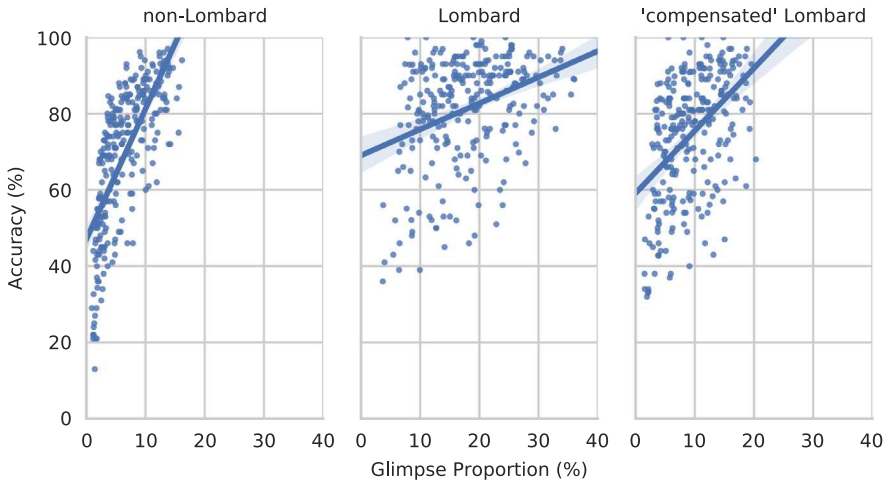

Fig. 6. Accuracy versus GP for different Lombard conditions. Each point represents the averages of an individual speaker at a particular SNR. The line represents the linear regression to the data points and the shaded area around the line the 95% confidence interval for the regression estimate.

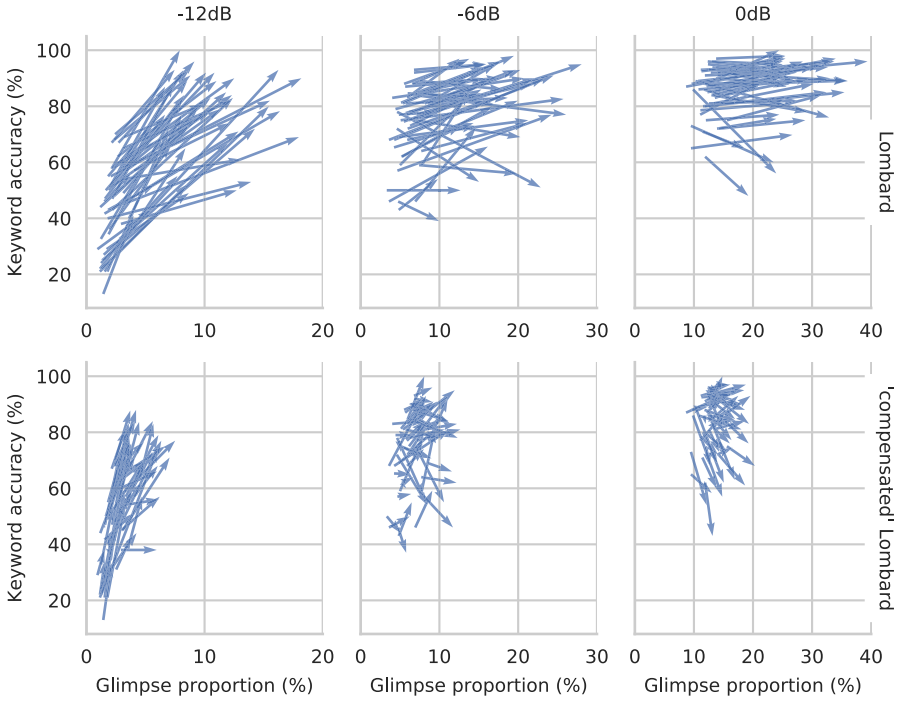

Fig. 7. Each arrow shows the shift in GP vs accuracy for a speaker when comparing normal speech to Lombard speech (top row) or 'compensated' Lombard speech (bottom row). Separate plots are shown for −12 dB (left), −6 dB (center) and 0 dB (right).

of the statistical model adaptation being less effective for outliers. In either case, it is evidence of a speaker variability that must be remembered before making general statements about the relative difficulty of automatically recognising normal versus Lombard speech in noise.

Finally, it should be acknowledged that the most extreme SNRs used in Study I and II are unlikely to be observed in real conversational settings. Normal conversational level lies at around 70 dB SPL (Moore, 2012) and would therefore require an 80 dB SPL noise background to achieve an SNR of −10 dB. At these background levels we have seen that we can expect Lombard gains of around 3–13 dB which would then result in speech being received at SNRs of −7 to +3 dB (depending on the speaker's Lombard effect). However, the choice of SNR settings has been motivated by the need to have the very low GP values that are necessary to make the small vocabulary Grid task sufficiently challenging. In a recognition task that was closer to meeting the demands of real applications, it would be expected that noise robustness problems and potential Lombard benefits would be observable at the higher SNRs more typical of everyday conversational environments.

6. Study III: the Lombard effect and visual speech recognition

6.1. Motivation

When communicating in noisy conditions, humans naturally exploit visual speech information concentrated in lip and jaw movements (Sumby and Pollack, 1954). The visual speech signal can greatly improve speech intelligibility in noise, in part because it helps release masking by driving auditory attention (Varghese et al., 2012) and in

part because it redundantly encodes phonetic information (Massaro et al., 1996). In fact, with sufficient contextual information, the visual signal alone can support accurate speech recognition – as demonstrated in visual-only lip-reading studies (e.g., Altieri et al., 2011) and in recent advances in automatic lip-reading using end-to-end systems (Assael et al., 2016).

There have been many previous studies of both human and machine audio-visual and visual-only speech recognition, however, few of these have employed video signals captured from *Lombard* speakers (notable exceptions include Vatikiotis-Bateson et al. (2007) and Davis et al. (2006)). This lack of study is particularly true for robust ASR where the non-availability of sufficient Lombard-style training data has been a serious obstacle. Instead, ASR studies have taken databases recorded in clean conditions (such as the AV Grid corpus (Cooke et al., 2006)) and artificially added noise. The one existing work that has examined audio-visual Lombard ASR has employed only 7 speakers, only 3 of which read full sentences (Heracleous et al., 2013). The continued neglect of Lombard visual speech in audio-visual ASR studies is particularly troublesome given that visual speech features are most likely to be useful to ASR in the low SNR situations that induce Lombard effects.

In this study we repeat the mismatched and matched recognition studies of the previous sections but this time using the visual component of the AV Lombard Grid corpus. Our first question is whether or not Lombard speech is sufficiently different in the visual domain that it produces mismatch effects in unadapted systems? The answer to this is not obvious, despite the strong mismatches observed in the audio data, considering that many of the major components of the Lombard reflex are modifications to the voicing source which will not have a direct impact on visual features. The second question is whether or not the visual Lombard speech signal can deliver higher speech recognition accuracy in well-adapted systems? Given that the visual domain is free of energetic masking, this question relates directly to the discussion of 'intrinsic intelligibility' that was introduced in the previous section.

6.2. Methods

6.2.1. Visual feature extraction

All systems have been trained on 2-D DCT transforms of the lip-region – similar to features that have performed well in the front-end of other speaker-dependent Grid corpus visual speech recognition systems (e.g., Zeiler et al., 2016). A three step process is used. First, full face tracking is employed to normalise out effects of head pose. Second, we locate a fixed size region of interest centred on the speaker's mouth. Finally, a 2-D DCT is used to perform dimensionality reduction and decorrelation. Although the AV Lombard Grid corpus is recorded with a camera fixed relative to the head (and hence free of head motion), to enable future comparisons, the feature extraction was designed to work equally well with the less controlled AV Grid corpus.

Each Lombard Grid frontal face video is decoded into a sequence of frames. A set of 68 landmark features is detected in each frame of the video using the ensemble regression tree method of Kazemi and Sullivan (2014) as implemented in the open-source *dlib* toolkit (King, 2009). In order to normalise the effects of head pose, we employ a set of template landmarks based on an average face looking directly at the camera. For each frame of the utterance, we calculate the affine transform that best maps the eyes and nose landmarks onto this template. This defines a sequence of transforms. We then select a single transform from this sequence that lies closest to the mean of the sequence. This single transform is then applied to every image frame in the video sequence. Using a single transform protects against gross misdetections in isolated frames, but it leaves a small amount of within-utterance head motion.

From the normalised images we then proceed to extract the lip region-of-interest (ROI). The ROI is based on the locations of the outer lip landmarks in the transformed space. For each frame we compute a bounding box around these landmarks. The width and height of the ROI is then set as the maximum width and maximum height over this set of bounding boxes multiplied by a factor of 1.2 (to capture the immediate context). The centre of each ROI in the frame sequence is set by linear regression through the sequences of bounding box centres – this linear fit captures the majority of previously-explained within utterance head-motion.

Note that although the size of the ROI remains fixed for each utterance it will vary between utterances – and to a larger extent between speakers. This variability is largely driven by speaker-camera distance and does not inform speech recognition, so it is therefore removed by rescaling each ROI to a fixed size of 240 by 240 pixels. Each pixel is then reduced from an RGB vector to a single scalar by conversion to the YUV colour space and then keeping the luminance value (Y). A 2-D DCT is applied to the matrix of luminance values. Finally, a 36 element feature vector is defined by concatenating the first eight diagonals of the DCT matrix.

6.2.2. Recognition systems

The design of the ASR systems remains the same as in the previous studies, and training follows the same recipes. Due to the low frame rate of the visual features (25 Hz), only their first order derivatives are used in the visual recognition system.

As in the case of the acoustic recognition system, the GMM-HMM models are initialized from a uniform alignment and trained in an alternating sequence of parameter estimation and alignment stages of increasing complexity. The recipe for training the system (implemented using Kaldi) is the same as the one used in the previous sections and includes LDA, MLLT and fMLLR SAT.

Separate systems have been trained for both non-Lombard (NL) visual speech and Lombard (L) visual speech. Performance has been measured for all speakers in both matched conditions (i.e., NL-NL, L-L) and mismatched conditions, i.e., training on NL and testing on L (NL-L).

6.3. Results

Results of the matched and mismatched recognition experiments are shown in Table 4. The table shows the keyword recognition accuracies averaged over all utterances and the standard errors. For the matched non-Lombard condition the performance is 41.3%. In the matched Lombard condition performance rises by nearly 10–51.0%. A single sided *t*-test on the per-speaker accuracy changes shows the increase in performance to be significant for both the mismatched and matched Lombard conditions ($t = 3.97$, $p < 0.001$ and $t = 8.08$, $p < 0.001$ respectively).

There have been previous suggestions that there are significant differences in clear speech styles of male and female speakers (Tang et al., 2015). We were therefore interested to see if the performance improvement was concentrated in one gender. The breakdown by gender shows that this is not the case: performance gains are statistically present for both male and female speakers. Although the effect is larger in female talkers (11% vs 7%) this difference is not significant (see Fig. 8).

The mismatch experiment demonstrates a surprising finding. Whereas we might expect the mismatched (NL-L) results to be poorer than the matched results (NL-NL), they are in fact significantly better.

Table 4
Keyword (digit & letter) recognition accuracy (and std. error) for matched and mismatched visual speech recognition systems.

	NL-NL	NL-L	L-L
Overall	41.3% (0.7%)	45.9% (0.7%)	51.0% (0.7%)
Female	42.0% (0.9%)	46.4% (0.9%)	53.7% (0.9%)
Male	40.4% (1.1%)	45.2% (1.1%)	47.4% (1.1%)

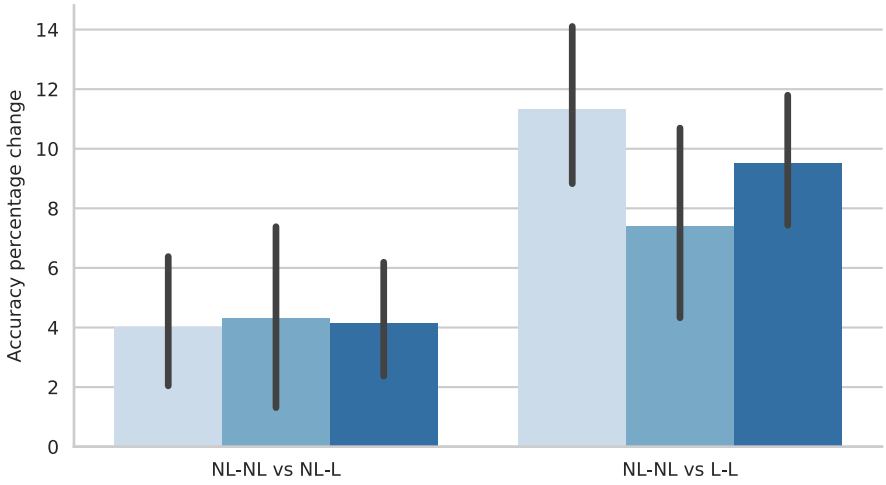

Fig. 8. Per-speaker average accuracy percentage change of the visual speech recognition systems from non-Lombard to Lombard (mismatch) and to Lombard (matched). From light to dark: female, male and overall. The height indicates the average across utterances while the black line depicts the standard error.

The improved discriminability of the visual Lombard features appears to more than compensate for the mismatch.

6.4. Discussion

The results of the current study (Fig. 8) show that visual Lombard speech can be recognised with significantly better accuracy than non-Lombard speech (+9%), even when being recognised with mismatched models trained on non-Lombard data (+4%). These increases are consistent across genders. This finding is in apparent contradiction with the earlier study of Heracleous et al. (2013) which observed a negative impact on recognition accuracy for visual Lombard speech.

Many factors may play a role in explaining the discrepancy between the two studies. In particular, Heracleous et al. (2013) contains two sub-studies one with only 4 speakers and the other with 3. Each study was based on different materials and different recognition systems. The large inter-speaker variability that has been observed in our work demonstrates the danger in drawing conclusions from such small sample sizes.

Our results indicate that Lombard speech is beneficial to automatic visual speech recognition even in the case where systems have not been exposed to utterances under such conditions. Analysis across speakers reveals that this effect is also consistent on a per-speaker basis.

7. Summary and conclusions

The paper has revisited the impact of Lombard speech on automatic speech recognition using a new audio-visual Lombard corpus. In particular, conclusions have been drawn from the analysis of 54 speakers – considerably more than have been used in the majority of previous studies. This scale has allowed us to arrive at significant findings despite the very large inter-speaker variabilities observed here and throughout the Lombard literature.

The first study examined the deleterious effects of Lombard mismatch in the audio domain. We considered a system that follows the typical noise robustness route of training on plain speech mixed with artificially added noise. It was seen that when such a system receives a Lombard utterance (i.e., typical of how the utterance would have genuinely been produced in noise), the resulting mismatch severely reduces performance. Much of this performance reduction can be recovered by simply correcting the Lombard gain (i.e., retraining with plain speech at the corrected SNR). However, even after appropriate gain normalisation, there remains a mismatch that, in our experiments, was sufficient to reduce recognition performance by an amount equivalent to a 3 dB reduction in SNR.

The second study took the opposite perspective and measured the potential performance benefit in a system well-adapted to Lombard speech versus one trained and tested on plain speech. Here it was seen that there was a clear and expected benefit from the reduction in SNR due to the Lombard gain. However, it was seen that a degree of Lombard benefit persisted even when the SNR is adjusted to that of the plain speech – a finding that has also been observed in human listening studies (Lu and Cooke, 2009). Analysis across speakers suggested that this benefit was most likely to be driven by a reduction in masking (i.e., due to redistribution of the speech energy). There was no evidence that performance benefits were due to an increased intrinsic intelligibility of the signal. On the contrary, at the highest SNRs, where masking effects are less significant there was seen to be a net *decrease* in performance.

The final study looked at the impact of the Lombard effect in visual speech recognition. Surprisingly, when training on non-Lombard speech, testing on mismatched Lombard speech produced a small (but significant) performance gain relative to testing on matched non-Lombard speech. This may be in part because the visual features that were used had an inbuilt normalisation for the degree of mouth opening. More importantly, however, when training and testing in matched conditions, the visual Lombard effect was seen to afford a large improvement in recognition performance (i.e., compared to a non-Lombard baseline). This is a novel finding that has not been reported in the literature, although it is in accordance with audio-visual speech *perception* studies which have found that the visual signal increases intelligibility more in Lombard speech than it does in plain speech (Vatikiotis-Bateson et al., 2007).

A key motivation for this work has been to examine the potential impacts of using artificial speech-plus-noise mixing in automatic speech recognition. Clearly, in trying to address this aspect of speech realism, the work has necessarily fallen short of realism in other respects. Particularly, the desire to perform a controlled comparison of Lombard versus non-Lombard speech has dictated the use of a read-speech task in simple, easily-repeatable noise conditions. Second, the practicality of collecting very large amounts of data in laboratory conditions has meant we have had to focus on a small vocabulary recognition task. For these reasons caution is needed when considering the *extent* to which the effects seen in these studies will apply in any particular real system. However, with this caveat aside, there appear to be two generally applicable conclusions: i) systems that are designed to operate in noise will benefit from being trained on well-matched Lombard speech data, ii) the results of speech recognition evaluations that employ artificial mixing need to be treated with caution. Whether such experiments under-estimate or over-estimate the difficulty of the true speech-in-noise problem will depend on the balance between the opposing effects

of reduced masking versus increased speaker variability seen in the Lombard condition.

Acknowledgements

This research has been funded by the UK Engineering and Physical Sciences Research Council (EPSRC project AV-COGHEAR, EP/M026981/1) and by the Saudi Ministry of Education, King Saud University.

References

Alghamdi, N., Maddock, S., Barker, J., Brown, G., Marxer, R., 2018. An audio-visual corpus for the study of Lombard speech. J. Acoust. Soc. Am. submitted.
Altieri, N.A., Pisoni, D.B., Townsend, J.T., 2011. Some normative data on lip-reading skills (l). J. Acoust. Soc. Am. 130 (1), 1–4.
Anastasakos, T., McDonough, J., Makhoul, J., 1997. Speaker adaptive training: A maximum likelihood approach to speaker normalization. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, pp. 1043–1046.
Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N., 2016. Lipnet: end-to-end sentence-level lipreading. arXiv:1611.01599.
Barker, J., Cooke, M., 2007. Modelling speaker intelligibility in noise. Speech Commun. 49 (5), 402–417.
Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2013. The PASCAL CHiME speech separation and recognition challenge. Comput. Speech Lang. 27 (3), 621–633. http://dx.doi.org/10.1016/j.csl.2012.10.004.
Bou-Ghazale, S.E., Hansen, J.H., 2000. A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE Trans. Speech Audio Process. 8 (4), 429–442.
Brown, G.J., Cooke, M., 1994. Computational auditory scene analysis. Comput. Speech Lang. 8 (4), 297–336.
Brumm, H., Zollinger, S.A., 2011. The evolution of the Lombard effect: 100 years of psychoacoustic research. Behaviour 148 (11–13), 1173–1198.
Cooke, M., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119 (3), 1562–1573.
Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. 120 (5), 2421–2424.
Cooke, M., Mayo, C., Villegas, J., 2014. The contribution of durational and spectral changes to the lombard speech intelligibility benefit. J. Acoust. Soc. Am. 135 (2), 874–883.
Davis, C., Sironic, A., Kim, J., 2006. Perceptual processing of audiovisual Lombard speech. Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. University of Auckland, New Zealand.
Fitzpatrick, M., Kim, J., Davis, C., 2013. Auditory and auditory-visual Lombard speech perception by younger and older adults. Proceeding of the International Conference on Auditory-Visual Speech Processing (AVSP).
Fitzpatrick, M., Kim, J., Davis, C., 2015. The effect of seeing the interlocutor on auditory and visual speech production in noise. Speech Commun. 74, 37–51.
Garnier, M., Henrich, N., 2014. Speaking in noise: how does the lombard effect improve acoustic contrasts between speech and ambient noise? Comput. Speech Lang. 28 (2), 580–597.
Glasberg, B.R., Moore, B.C., 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 47 (1), 103–138.
Hansen, J.H., 1989. Analysis and compensation of stressed and noisy speech with application to robust automatic recognition. Signal Process. 17 (3), 282.
Hansen, J.H., Bou-Ghazale, S.E., 1995. Robust speech recognition training via duration and spectral-based stress token generation. IEEE Trans. Speech Audio Process. 3 (5), 415–421.
Hansen, J.H., Clements, M.A., 1995. Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress. IEEE Trans. Speech Audio Process. 3 (5), 407–415.
Hansen, J.H., Varadarajan, V., 2009. Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition. IEEE Trans. Audio Speech Lang. Process. 17 (2), 366–378.
Heracleous, P., Ishi, C.T., Sato, M., Ishiguro, H., Hagita, N., 2013. Analysis of the visual Lombard effect and automatic recognition experiments. Comput. Speech Lang. 27 (1), 288–300.
Hirsch, H.-G., Pearce, D., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP). 4. pp. 29–32.
Huang, F.J., Chen, T., 2001. Consideration of Lombard effect for speechreading. Proceedings of the 2001 IEEE Fourth Workshop on Multimedia Signal Processing. IEEE, pp. 613–618.
Junqua, J.-C., 1993. The Lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am. 93 (1), 510–524.
Kazemi, V., Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1867–1874.
Kim, J., Sironic, A., Davis, C., 2011. Hearing speech in noise: seeing a loud talker is better. Perception 40 (7), 853–862.
King, D.E., 2009. Dlib-ml: A machine learning toolkit. J. Mach. Learn. Res. 10 (July), 1755–1758.
Lombard, E., 1911. Le signe de l'elevation de la voix. Ann. Maladies Oreille, Larynx, Nez, Pharynx 37, 101–119.
Lu, Y., Cooke, M., 2008. Speech production modifications produced by competing talkers, babble, and stationary noise. J. Acoust. Soc. Am. 124 (5), 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. 51 (12), 1253–1262.
Massaro, D.W., Cohen, M.M., Smeele, P.M., 1996. Perception of asynchronous and conflicting visual and auditory speech. J. Acoust. Soc. Am. 100 (3), 1777–1786.
Meutzner, H., Ma, N., Nickel, R., Schymura, C., Kolossa, D., 2017. Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates. Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5320–5324. http://dx.doi.org/10.1109/ICASSP.2017.7953172.
Moore, B.C., 2012. An Introduction to the Psychology of Hearing. Brill.
Parihar, N., Picone, J., Pearce, D., Hirsch, H.G., 2004. Performance analysis of the Aurora large vocabulary baseline system. Proceedings of the 2004 European Signal Processing Conference (EUSIPCO). Vienna, Austria. pp. 553—556.
Patel, R., Schell, K.W., 2008. The influence of linguistic content on the Lombard effect. J. Speech, Lang. Hear. Res. 51 (1), 209–220.
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 (4), 434–446.
Pittman, A.L., Wiley, T.L., 2001. Recognition of speech produced in noise. J. Speech, Lang. Hear. Res. 44 (3), 487–496.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The Kaldi speech recognition toolkit. Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE Signal Processing Society.
Stanton, B.J., Jamieson, L., Allen, G., 1988. Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, pp. 331–334.
Stanton, B.J., Jamieson, L., Allen, G.D., 1989. Robust recognition of loud and Lombard speech in the fighter cockpit environment. Proceedings of the 1989 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, pp. 675–678.
Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26 (2), 212–215.
Summers, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., Stokes, M.A., 1988. Effects of noise on speech production: acoustic and perceptual analyses. J. Acoust. Soc. Am. 84 (3), 917–928.
Svirsky, M.A., Lane, H., Perkell, J.S., Wozniak, J., 1992. Effects of short-term auditory deprivation on speech production in adult cochlear implant users. J. Acoust. Soc. Am. 92 (3), 1284–1300.
Tang, L.Y., Hannah, B., Jongman, A., Sereno, J., Wang, Y., Hamarneh, G., 2015. Examining visible articulatory features in clear and plain speech. Speech Commun. 75, 1–13.
Tartter, V.C., Gomes, H., Litwin, E., 1993. Some acoustic effects of listening to noise on speech production. J. Acoust. Soc. Am. 94 (4), 2437–2440.
Varghese, L.A., Ozmeral, E.J., Best, V., Shinn-Cunningham, B.G., 2012. How visual cues for when to listen aid selective auditory attention. J. Assoc. Res. Otolaryngol. 13 (3), 359–368.
Vatikiotis-Bateson, E., Barbosa, A.V., Chow, C.Y., Oberg, M., Tan, J., Yehia, H.C., 2007. Audiovisual Lombard speech: reconciling production and perception. Proceeding of the International Conference on Auditory-Visual Speech Processing (AVSP). pp. 41.
Vatikiotis-Bateson, E., Chung, V., Lutz, K., Mirante, N., Otten, J., Tan, J., 2006. Auditory, but perhaps not visual, processing of Lombard speech. J. Acoust. Soc. Am. 119 (5). 3444–3444
Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M., 2013. The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 126–130.
Wakao, A., Takeda, K., Itakura, F., 1996. Variability of Lombard effects under different noise conditions. Proceedings of the Fourth International Conference on Spoken Language Procesing (ICSLP). 4. IEEE, pp. 2009–2012.
Young, S.J., Odell, J.J., Woodland, P.C., 1994. Tree-based state tying for high accuracy acoustic modelling. Proceedings of the Workshop on Human Language Technology. Association for Computational Linguistics, pp. 307–312.
Zeiler, S., Nicheli, R., Ma, N., Brown, G.J., Kolossa, D., 2016. Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis. Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2797–2801.