Analysis of Lombard effect under different types and levels of noise with application to In-set Speaker ID systems

Vaishnevi S. Varadarajan, John H.L. Hansen

Center for Robust Speech systems
Department of Electrical Engineering
University of Texas at Dallas, USA
varadara@colorado.edu, john.hansen@utdallas.edu

Abstract[1]

This paper presents an analysis of Lombard speech produced under different types and levels of noise. The speech used for the analysis forms a part of the UT-SCOPE database and consists of sentences from the well-known TIMIT corpus, spoken in the presence of highway, large crowd and pink noise. Differences are shown to exist in the speech characteristics under these varying noise types. The deterioration of the EER of an in-set speaker identification system trained on neutral and tested with Lombard speech is also illustrated. A clear demarcation between the effect of noise and Lombard effect on noise is also given by testing with noisy Lombard speech. The effect of test-token duration on system performance under the Lombard condition is addressed. It is seen that test duration has no effect on the EER under Lombard effect. The average EER for 3s test duration is 14.7, 28.3, 48.2, 51.3 for neutral clean, clean Lombard, noisy neutral and noisy Lombard respectively, and 7.2, 26.4, 45.8, 50.8 respectively for 12s test duration.

Index Terms: Lombard effect, In-set speaker recognition, Speaker ID

1. Introduction

The advances in speech technology have led to an increased deployment of automatic speech systems in varying environments such as factories, busy offices, cars, lecture halls, wireless PDAs, etc. This presents challenges to researchers, not only in dealing with the environmental noise per se, but also with the changes in the speech characteristics that arise from the well-known Lombard effect. The Lombard effect may be defined as speech produced due to increased vocal effort on the part of the speaker to improve the communication efficiency over environmental noise. It has been shown by Rajasekaran et al. [1] that Lombard effect degrades speech system performance to a greater degree than noise itself. In [3], speaker recognition performance for several stressful conditions including Lombard effect was considered. Several approaches in the past have aimed at bridging the differences between training and testing conditions with respect to speech recognition systems. One of the earliest compensation schemes developed by Chen Y [9] performs a hypothesis driven cepstral domain compensation on stressed speech. In another scheme [10], a slope-dependent weighting of the metric was performed to account for the differences in the spectral slope for neutral and Lombard speech. A linear transformation of LPC cepstral features was suggested in [11] with applications to a DTW based speech recognition system. Yet another scheme [14] used the source-generator framework for compensation of speaker stress. The compensation was an additive component to the MFCC vector, and this factor depended on the nature of speech frame (voiced, unvoiced, transitional). The compensation factor was uniform for all levels of noise.

Many studies have been initiated to analyze the characteristics of speech produced in noise ([2], [4], [5], [6]). These analyses have considered only individual words spoken under the Lombard effect whereas real speech systems use sentences as test utterances. In this paper, we perform our analyses on sentences. Further, previous studies have concentrated on speech produced under a single noise type. Since performance of speech systems vary according to noise type, one can expect the same for the Lombard effect also. Hence, compensation schemes for the Lombard effect should inherently depend on the nature of environmental noise. To our best knowledge, this is the first paper to perform a scientific study of the Lombard effect under varying noise levels and types with specific applications to speaker ID systems.

The paper is organized as follows. Sec. 2 of the paper describes the speech database used in the analysis. The next section details the various analyses performed on the Lombard speech. Differences from the past studies are also discussed. In sec. 4, in-set speaker ID tests and results are presented. We close in sec. 5 with some concluding remarks.

2. The UT-SCOPE database

The speech data used for this study forms a part of the UT-SCOPE (Speech under Cognitive and Physical stress and Emotion) database, details of which can be found in [7]. Lombard speech was produced under three different noise types: highway noise (HWY) in a car traveling at 65 mph with windows half open, large crowd noise (LCR), and pink noise (PNK). The noise, calibrated using a Quest sound level meter was played at different levels. The levels used were 70, 80 and 90 dB-SPL for highway and large crowd noise; 65, 75 and 85 dB-SPL for pink noise. Open-air headphones were used to play the noise to enable human feedback of the speech produced, back to the subject's ears. A pure tone hearing test was also conducted on every subject to check for hearing loss, if any. The speech was recorded in two sessions using three different microphones - a throat microphone, a close talking and a far field microphone. All recordings were made in an ASHA certified, double walled sound booth using a multi-track

1This work was supported by a Grant from the U.S. Air Force under contract No. FA8750-05-C-0029, RADC under contract No. FA8750-05-C-0029, and by the Univ. of Texas at Dallas under Project EMMITT.

FOSTEX DAT recorder with gain adjustments for individual channels. 35 subjects (15 males and 20 females) participated in the data collection.

The speech consists of 100 sentences from the TIMIT database for the neutral condition and 20 sentences for each of the Lombard effect condition. These 20 sentences form a subset of the ones used for the neutral condition. Each of the 10 conditions also includes 5 tokens each of the 10 digits (0-9) and spontaneous speech of one-minute duration.

3. Analysis of Lombard speech

Speech from 7 male speakers, produced at the close-talking microphone, was used in the following analyses. The Lombard speech was analyzed for duration, silence duration, frame energy distribution and spectral tilt. The results of the study are discussed below.

3.1. Sentence duration

Duration analysis of single words in [2], [4], [5] and [6] reveals that the duration of vowels increases and that of the stops and fricatives decreases and hence average word duration increases under Lombard effect. For this study, duration of the twenty sentences in each Lombard effect condition was normalized by the corresponding duration under neutral condition. This was done to remove the effect of sentence length, which varied within the chosen set. It was found that on average, the sentence duration decreases. This could be because of two reasons:

- Both silence and word durations decrease,
- Silence duration decreases more than the increase in word duration

To clarify this, a frame-energy based approach was used to eliminate silence frames from the sentences. Frames of length 20ms with an overlap of 10ms were used and those above a certain threshold were selected as speech frames. From the frame count, the speech duration was computed. Normalized durations were calculated for the 20 sentences for the 7 male speakers and were fit to a Gaussian distribution. The mean and the variance of the distribution are shown in Table 1 for each of the nine conditions (i.e. 3 noise types, each at 3 levels).

Table 1. *Mean and variance of the Gaussian pdf modeling sentence duration. Noise level 1, 2, 3 are [LCR, HWY:70,80,90 dB-SPL, PNK:65,75,85 dB-SPL]*

Noise Type	Noise Level 1		Noise Level 2		Noise Level 3	
	mean	var	mean	var	mean	var
HWY	0.998	0.18	0.987	0.14	0.987	0.15
LCR	0.954	0.21	0.955	0.15	0.97	0.21
PNK	0.933	0.17	0.915	0.16	0.945	0.15

From the values above it is clear that the average duration of a sentence decreases under Lombard effect. This result is quite contrary to the duration results for individual words [4].

3.2. Duration of silence

A study of the duration of silence in the individual utterances revealed some interesting trends. As described above, a frame-energy based scheme was used to detect frames of silence. The percentage of silence in each of the utterances was thus computed. The average silence percentages are shown in Fig. 1, where duration of silence decreases under Lombard effect. This implies a sense of urgency on the part of the speaker, which occurs due to the persistent exposure of the environmental noise to the speaker's ears. Also, it can be seen that this duration depends on the noise level only for highway noise and is quite consistent for large crowd and pink noise across different noise levels.

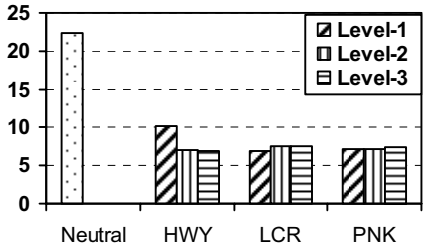

Fig 1. *Percentage of silence across 3 noise types and levels. Noise levels 1,2,3 are [LCR, HWY:70,80,90 dB-SPL, PNK:65,75,85 dB-SPL]*

3.3. Energy distribution

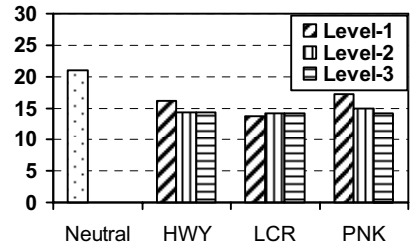

Fig 2. *Percentage of low energy frames in speech*

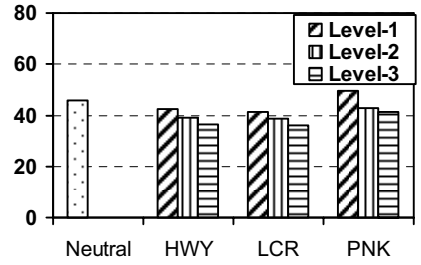

Fig 3. *Percentage of middle energy frames in speech*

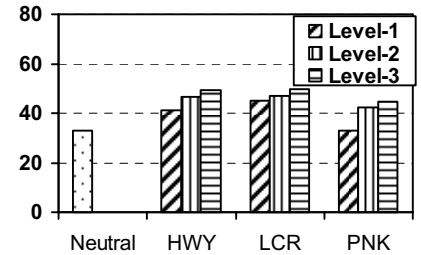

Fig 4. *Percentage of high energy frames in speech*

Histograms of frame energy distribution for all the utterances were used to compute percentages of low, middle and high-energy frames for the 7 speakers under all the ten conditions (i.e. neutral clean, and Lombard effect for 9 noise exposures), the results of which are illustrated in Fig. 2, 3 and 4. From the

figures, it is evident that under noise, there is a migration from low and middle energy regions to high-energy zones. Also, the reduction in percentage for low energy frames is quite sharp in going from neutral to noisy condition as compared to changing between noise levels. Isolated words, on the other hand, demonstrated migration of energy from low and high to middle-energy regions for vowels and from low to middle-energy regions for consonants.

3.4. Spectral tilt

Variations in glottal spectral tilt were investigated for the 7 male speakers. The approach for estimating the glottal spectral tilt is detailed in [4]. Forced alignment of the speaker utterances were performed at the phone level and nasals were removed. The speech frames above a certain threshold were chosen and the periodogram computed. The resulting periodograms were averaged and a linear regression was performed to compute the slope of the glottal spectrum. Table 2 gives the average spectral slope for the different Lombard effect conditions.

Table 2. *Mean spectral tilt for different Lombard effect condition. Noise levels 1,2,3 are [LCR, HWY:70,80,90 dB-SPL, PNK:65,75,85 dB-SPL]*

Noise Type	**Noise Level 1** (dB/octave)	**Noise Level 2** (dB/octave)	**Noise Level 3** (dB/octave)
NEU	-15.92		
HWY	-15.3	-14.6	-13.9
LCR	-15.6	-14.5	-13.8
PNK	-15.4	-15.0	-14.1

The table above indicates that the spectral slope decreases under Lombard effect, implying that the energy in the high frequency of the glottal spectrum increases. These findings are consistent with those for isolated words [4].

4. In-set speaker ID performance

4.1. System details

The speech data collected under the different Lombard effect conditions were tested on an in-set speaker ID system. An in-set speaker ID system is one that identifies if the speech input belongs to one of the group of speakers defined in the system. It only classifies the speech as in-set or out-of-set but does not identify the speaker from the speech. This back-end system employs a Universal Background Model (UBM) constructed from a selected set of speakers. A speaker specific MAP adapted Gaussian Mixture Model (GMM) is obtained from the UBM for each of the trained in-set speakers. The scores obtained by comparing the test utterances with the trained speaker models were normalized and thresholds were set using unconstrained cohort normalization likelihood ratio testing [8]. Further details of the GMM-UBM system can be found in [9]. Equal error rates (EER) were obtained using the in-set speaker ID system for the different Lombard effect conditions.

4.2. Experimental setup

4.2.1. Speaker and development set

A set of 30 speakers was chosen. The population consisted of 19 females and 11 males. 15 were in-set and 15 out-of-set. Out of the 15 in-set speakers, 9 were females and 6 males. There were 10 female and 5 male out-of-set speakers. The development set for the UBM consisted of 60 speakers chosen from the TIMIT corpus. The male-female ratio in the development set was maintained the same as the in-set speakers.

4.2.2. Front-end processing

Speech from all speakers was windowed with a Hamming window of 20ms duration with 10ms overlap rate. A 23-dimensional feature vector consisting of 19-dimensional MFCC's and 4 spectral center of gravity coefficients was extracted from all the speech data [9].

4.3. Experiments and results

Two sets of experiments were performed on the speaker ID system. Both experiments used training data consisting of ~30s (10 sentences) of neutral speech. The first set of tests investigated the degradation caused by Lombard effect only. The neutral-trained speaker ID system was tested with clean neutral and noise-free Lombard speech. The effect of test utterance duration was also investigated by using two sets of test utterance length, 3s and 12s. The results are shown in Tables 3 and 4.

Table 3. *EER (%) of in-set speaker ID system using 3s clean test utterances. EER with Neutral speech=14.67%*

Noise Type	**Noise Level 1**	**Noise Level 2**	**Noise Level 3**
HWY	23.16	32.67	34.83
LCR	25.83	29.5	30.33
PNK	22.17	25	31.5

Table 4. *EER (%) of in-set speaker ID system using 12s clean test utterances. EER with Neutral speech=7.2%*

Noise Type	**Noise Level 1**	**Noise Level 2**	**Noise Level 3**
HWY	20	29.5	34
LCR	24.5	30.17	28.83
PNK	16.8	22.16	31.5

From the above results, one can clearly see that Lombard speech degrades the performance of a speaker ID system. The average increase in the EER for under the different Lombard conditions with 3s test tokens is 93%, relative to the EER under neutral condition and that in the 12s test case is 266%. From the absolute values of the EER, we can see that an increase in the test duration helps in improving the EER under the neutral condition by about 50%, but the average improvement under the Lombard conditions is only 7.68%. Also, we see that with increased test duration, EER reduction under the highest level (Level 3) of noise is negligible (2.4 %). Hence, increase in the test duration does not improve the EER under Lombard

conditions and therefore, the Lombard effect changes the spectral structure to the point where additional test material cannot recover the performance.

The second set of experiments was performed by degrading the neutral and Lombard speech test tokens. The training was however done with clean neutral speech only. This was considered in order to determine if speaker ID performance is more significantly impacted by noise type/level, or speech production changes due to Lombard effect. The noise used for producing the Lombard conditions was used for degrading the respective utterances. The SNRs used for large crowd, highway and pink noise conditions were 5dB, -5dB and 0 dB respectively. These noise levels represent the exact noise present when collecting noise-free Lombard speech. The experiments were repeated for 3s and 12s test utterances. The results are summarized in Tables 5 and 6.

Table 5. *EER (%) of in-set speaker ID system using degraded 3s test utterances. Clean Neutral EER:14.67 %*

Noise Type	NOISY NEU	Noise Level 1	Noise Level 2	Noise Level 3
HWY	49.33	53.33	54.167	54.167
LCR	46.33	48.33	53	51.5
PNK	48.83	48.99	48	50.167

Table 6. *EER (%) of in-set speaker ID system using degraded 12s test utterances. Clean Neutral EER: 7.2%*

Noise Type	NOISY NEU	Noise Level 1	Noise Level 2	Noise Level 3
HWY	45.49	51.33	52.67	56.17
LCR	42.5	49.5	49.3	50
PNK	49.33	44.67	52.16	51.5

The first column in the above tables marked NOISY NEU represent the EER for neutral speech degraded with the respective noise types (i.e. noisy speech without the Lombard effect). It is evident from that the Lombard effect along with noise degrades system perform more than noise only. Also, we can see that the error rates are not additive, in the sense that the EER with noisy neutral and clean Lombard speech do not sum up to the EER with noisy Lombard speech.

Here, it is noticeable that the increase in the test duration does not help at high Lombard levels (Level 3). Lombard speech with noise clearly gives very poor performance. When speech enhancement algorithms for noisy Lombard speech do not address Lombard effect, we can only move up to the performance shown for clean Lombard speech in Tables 4 and 5. Achieving true effective performance for speech enhancement in noisy Lombard speech therefore requires normalization of the Lombard effect [13].

5. Conclusions and future work

This study has considered an analysis of the characteristics of speech under different types of Lombard conditions. It has been shown that the percentage of silence in speech is reduced significantly. Also, there is a migration of energy from the lower and middle to the higher levels due to noise. This migration is somewhat consistent across noise types. Also, the duration of sentences is reduced under Lombard effect. The reduction in duration depends on the type and level of noise. For certain noise types, the reduction is not significant between different noise levels. The spectral tilt of Lombard speech decreases, indicating an increase in the high frequency components under noise. EER of in-set speaker ID system shows degradation under Lombard effect. It has also been shown that increasing test duration does not improve the system performance under the Lombard conditions, indicating fundamental changes in phoneme spectral structure. Also, compensation for only noise under noisy Lombard conditions keeps the system performance far from baseline performance. This represents the first study to investigate the change in speech production for Lombard effect under different noise types and levels. It has also been shown that while noise impacts speaker ID performance, speech production under Lombard effect causes fundamental changes in spectral structure for a GMM that cannot be overcome by simply using longer test sequences. Earlier studies that have considered Lombard effect compensation for isolated-word recognition [10],[11],[12],[13],[14] could be considered to compensate the varying types of Lombard effect speech seen for the first time in this study.

6. References

[1] Rajasekaran P., Doddington G., Picone J. "Recognition of speech under stress and in noise", ICASSP 86, pg. 733-6.

[2] Summers W. et al. "Effects of noise on speech production: Acoustical and perceptual analyses", JASA, vol. 84, 1988.

[3] H.J.M. Steeneken, J.H.L. Hansen, "Speech Under Stress Conditions: Overview of the Effect of Speech Production on Speech System Performance," ICASSP 99.

[4] Hansen J.H.L., "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition", Thesis, Georgia Inst. Tech. July 1988.

[5] Stanton B.J. et al. "Acoustic-Phonetic analysis of loud and Lombard speech in simulated cockpit conditions", ICASSP 88

[6] Junqua J. "The Lombard reflex and its role on human listeners and automatic speech recognizers", J. Acoust. Soc. Amer., Jan. 1993.

[7] Varadarajan V, Hansen J.H.L, Ikeno A " UT-SCOPE – A corpus for Speech under Cognitive/Physical task Stress and Emotion", LREC workshop on speech under emotion, May 2006.

[8] Fortuna J., Sivakumaran P. et al. "Open-set speaker identification using adapted Gaussian mixture models", Interspeech 2005.

[9] Angkititrakul et al., "Cluster-dependent Modeling and Confidence Measure Processing", ICSLP 2004.

[10] Chen Y, "Cepstral domain talker stress compensation for robust speech recognition", IEEE Trans. ASSP, vol.36, April 1988.

[11] Stanton B.J. et al. "Robust recognition of loud and Lombard speech in the fighter cockpit environment", ICASSP 89.

[12] Takeda K. et al., "Variability of Lombard effects under different noise conditions", ICSLP 96, pg.2009-2012.

[13] Hansen J.H.L. et al. "Constrained Iterative Speech Enhancement with Application to Speech Recognition", IEEE. Trans. Sig. Proc. Apr. 1991.

[14] Hansen J.H.L. "MCE-ACC for Speech Recognition in Noise and Lombard Effect", IEEE Trans. SAP, vol.2, Oct.1994.