# Unsupervised Acoustic Analyses of Normal and Lombard Speech, with Spectral Envelope Transformation to Improve Intelligibility

*Elizabeth Godoy, Yannis Stylianou*

Institute of Computer Science, Foundation of Research and Technology Hellas (FORTH), Crete, Greece
egodoy@ics.forth.gr, styliano@ics.forth.gr

## Abstract

The "Lombard effect" describes how humans modify their speech in noisy environments to make it more intelligible. The present work analyzes Normal and Lombard speech from multiple speakers in an unsupervised context, using meaningful acoustic criteria for speech classification (according to voicing and stationarity) and evaluation (using loudness and intelligibility). These acoustic analyses using generalized classes offer alternative and informative interpretations of the Lombard effect. For example, the Lombard increase in intelligibility is shown to be isolated primarily to voiced speech. Also, while transients are shown to be less intelligible overall, the Lombard effect does not appear to distinguish between stationary and transient speech. In addition to these analyses, following recently published results illustrating that Lombard spectral modifications account for the largest increases in intelligibility, this work also examines spectral envelope transformation to improve speech intelligibility. In particular, speaker-dependent Normal-to-Lombard correction filters are estimated and, when applied in transformation, shown to yield higher overall objective intelligibility than Normal, and even Lombard, speech.

**Index Terms**: Lombard effect, spectral transformation

## 1. Introduction

When speaking in noisy environments, humans modify their speech in order to make it more intelligible. This phenomenon and its corresponding speech modifications are collectively referred to as the "Lombard effect" [1]. The Lombard effect is typically associated with an increase in pitch, a slower speaking rate, and a decrease in spectral tilt or spectral "flattening," among other changes [2, 3, 4]. In an effort to better understand the Lombard speech modifications as well as assess their influence on intelligibility, acoustic features are often examined jointly with contextual information, for example, using phonetic segmentation and labeling [3, 4] or articulatory features [5].

The present work analyzes the Lombard effect in an unsupervised context, i.e. without contextual information, using generalized acoustic classes and examining objective metrics evaluating speech intelligibility and loudness. In particular, an average loudness metric and an extended Speech Intelligibility Index (SII), both based on established standards for speech perception, are examined for the Normal and Lombard speech of multiple male and female speakers. Classification of frames is then considered based on voicing and stationarity. In this way, the Lombard effect analyses are interpreted purely in terms of general, yet meaningful, acoustic criteria, without need for corpus segmentation and labeling. Much like supervised contextual classification, these acoustic indicators separating voiced speech (vowels, nasals, etc) from unvoiced speech (certain consonants, plosives, fricatives, etc), and stationary (stable) speech from transitions between acoustic events, can also offer valuable insight on the Lombard effect. For example, work in [6, 7] suggests that modifications of transient parts of speech significantly impact intelligibility; the analyses in this work are able to directly examine the intelligibility of transients in addition to assessing their role in relation to the Lombard effect.

In addition to understanding and explaining the Lombard effect, there is currently interest in modifying speech in order to make it more intelligible; for example, in order to make speech synthesizers more understandable for listeners in noisy environments [8, 9]. In order to accomplish this goal, transformation of the spectral envelope is key, as the spectral envelope modifications associated with the Lombard effect have been shown to be most responsible for increasing speech intelligibility [10]. Accordingly, the present work examines spectral envelope transformation by applying a correction filter, inspired by the amplitude scaling proposed for voice conversion in [11], to Normal speech in order to render it more intelligible. An enhanced version of the correction filter, estimated using only the Lombard speech with the highest objective intelligibility (i.e., SII), is also evaluated.

The structure of this article is as follows. Section 2 first describes the speech corpora and estimation of acoustic parameters and criteria for evaluation. The differences in intelligibility and loudness between Normal and Lombard speech are then analyzed in terms of the generalized acoustic classes based on voicing and stationarity. Section 3 then describes and evaluates the proposed Normal-to-Lombard spectral envelope transformation to increase speech intelligibility. Finally, section 4 concludes and discusses future work.

## 2. Acoustic Analyses of Normal and Lombard Speech

### 2.1. Speech Data

The speech data is from the GRID corpora described in [3, 10] and includes 8 British English speakers (4 male, 4 female). There are 50 sentences per speaker and each sentence is recorded both in quiet conditions (Normal) and in speech-shaped noise at a 96dB level (Lombard). The speech sampling rate is 16kHz, downsampled from 25kHz, and the Lombard speech is aligned in time to the Normal speech using the VocALign software (www.synchroarts.com).

The speech analysis and synthesis are pitch-asynchronous (using a 30ms Hanning window and a 10ms step) and DFT-based with an FFT length of 2048. The synthesis is overlap-add and the frame reconstruction uses the DFT magnitudes (modified in the case of transformation) and phases from analysis (original phases).

### 2.2. Acoustic Parameters

Two acoustic indicators are used in the present work to classify speech according to voicing and stationarity, respectively. First, the voicing parameter is estimated as the number of zero-crossings in a frame, normalized by the frame length. After examining the voicing parameter distribution for all frames in the corpora, a threshold of 0.35 was selected in order to isolate the observed voiced (≈85% of frames) and unvoiced (≈15% of frames) modes in the speech. Second, the stationarity parameter is described in [12] and is estimated as the transition rate of the event functions capturing spectral envelope variations in a temporal decomposition of speech. Like the voicing analysis, after examining the stationarity parameter distributions, a threshold of 0.25 was selected in order to isolate stationary or stable (≈86% of frames) speech from transients or transitions.

For evaluations, two objective metrics are examined, based on established PEAQ and ANSI standards respectively related to perceived "loudness" and speech intelligibility. First, as Lombard speech is often described as "louder" than Normal speech (even when both have equal energy), the loudness is estimated, following the PEAQ standard [13]. Present evaluations then focus on the frequency range between 500-4500Hz, representing an inclusive formant region. The average loudness in this region, hereafter referred to as "Loudness," more closely relates to sensitivities in human speech perception than examining the entire range of spectrum frequencies. In addition to Loudness, the extended SII described in [14] is examined. The SII is calculated for the speech with white noise and babble noise added at -5dB and -10dB SNR, respectively. After observing the similar behavior of the SII metric for these noise conditions, their average SII was calculated and is presented in this work. This SII is also directly compared with the Loudness, as perceived loudness is expected to be correlated with intelligibility (i.e. generally speaking, louder speech can be seen to some extent as being more intelligible).

### 2.3. Loudness and SII

Figure 1 shows histograms of the Normal and Lombard Loudness and SII for all speakers and all frames in the corpora. Given the observed similarities in the Loudness and SII distributions, as well as the intuitive link between perceived loudness and intelligibility, the Pearson correlation coefficient for Loudness-SII was calculated and found to be 0.88 for Normal and 0.92 for Lombard speech, confirming high correlation between the two metrics. Thus, the metrics can be used interchangeably in analyses, though it should be noted that, while the SII depends on noise, the Loudness depends only on the speech signal. Examining Figure 1, there is a larger percentage of the Lombard speech that lies above 3 for Loudness and 0.6 for SII. These regions can be interpreted as those in which the "Lombard effect" is most present, that is, there are more Lombard frames perceived as louder and highly intelligible. Also, note that for the Lombard histograms, two modes are observed, with one that is less loud/intelligible than the other. These modes will be explained through the following analyses.

In order to better understand the histogram behaviors for Normal and Lombard speech, Table 1 provides the median values for Loudness and SII, also considering voiced-unvoiced and stationary-transient classes. In Table 1, it can be clearly seen that unvoiced and transient frames are less loud/less intelligible for both Normal and Lombard speech. However, upon close examination of the difference between the median Loudness and SII for voiced-unvoiced frames, it is evident that this difference is more substantial for the Lombard speech (with over 50% reduction for unvoiced frames compared to all frames). Thus, the two modes evident in Figure 1 for the Lombard speech can be mainly attributed to the voiced-unvoiced class differences: this is shown explicitly in Figure 2.

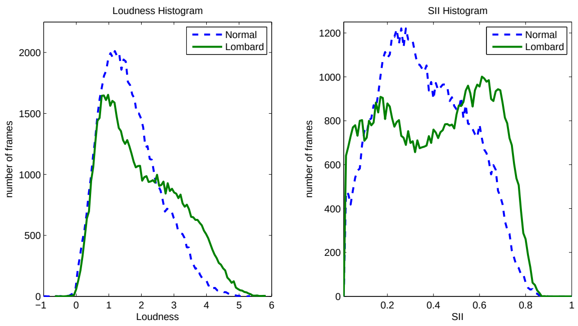

Figure 1: *Histograms of Loudness (left) & SII (right) for Normal & Lombard speech.*

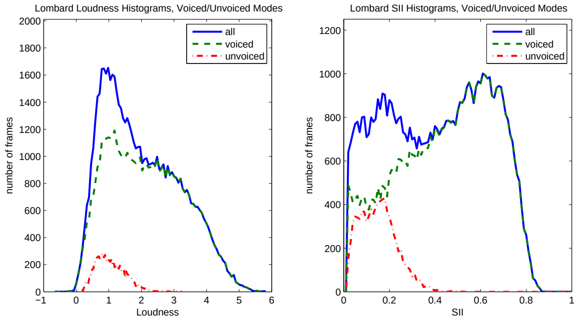

Figure 2: *Histograms of Loudness (left) & SII (right) for Lombard speech, with voiced-unvoiced mode separation.*

Table 1: *Median Loudness & SII for Voiced(V)-Unvoiced(UV) and Stationary(St.)-Transient(Tr.) classes.*

| | **All** | **V** | **UV** | **St.** | **Tr.** |
|---|---|---|---|---|---|
| Loud., Normal | 1.52 | 1.57 | 1.36 | 1.55 | 1.43 |
| Loud., Lombard | 1.81 | 2.08 | 1.00 | 1.95 | 1.40 |
| SII, Normal | 0.34 | 0.37 | 0.22 | 0.35 | 0.27 |
| SII, Lombard | 0.41 | 0.47 | 0.15 | 0.45 | 0.26 |

### 2.4. Lombard Effect: Influence of Voicing & Stationarity

In an effort to explicitly isolate the Lombard effect, the following analyses directly examine the difference in Loudness between the Normal and Lombard speech for each frame, exploiting the fact that the corpora are aligned in time. Figure 3 shows histograms of the average of this Lombard-Normal loudness difference in the 500-4500Hz region, including distributions for voiced-unvoiced and stationary-transient classes. In the left plot of Figure 3, it is shown that, while voiced Lombard speech is louder than that of Normal speech, unvoiced Lombard speech is actually less loud. In other words, the Lombard effect of increasing loudness or intelligibilitiy, is isolated to voiced speech. Moreover, this observation supports the findings in [4] that the Lombard effect involves a shift in energy from consonants to

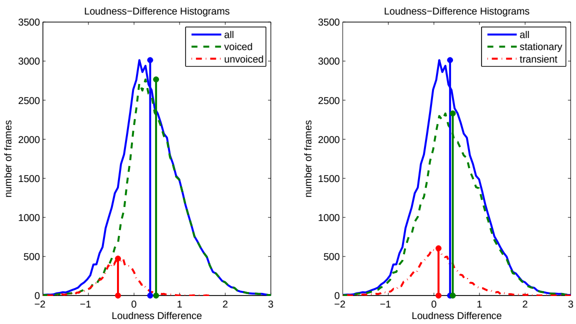

Figure 3: *Histograms of the Loudness Difference between Lombard & Normal speech, with frame separation based on voicing (left) & stationarity (right). Mean locations for each distribution are indicated with '*'.*

vowels. Alternatively for stationarity, in the right plot of Figure 3, there is no observed Lombard effect difference between stationary and transient speech. Thus, while Table 1 shows that transient speech is less loud/intelligible, Figure 3 suggests that transient speech contributes little to Lombard-Normal differences; in other words, the Lombard effect does not appear to differentiate between transient and stationary speech.

## 3. Spectral Envelope Transformation

In addition to the unsupervised acoustic analyses of Normal and Lombard speech, the present work also examines spectral envelope transformation to improve speech intelligibility. Similar to the idea in [10], the Normal speech is modified here so that the average spectral envelope resembles that of Lombard speech. Specifically, an approach is adopted that is inspired by the amplitude scaling proposed in [11] for voice conversion. For each speaker, the following process is carried out. First, a spectral envelope correction filter is estimated as the log-difference between the average Lombard and average Normal spectral envelopes, calculated using all frames. Note here that the spectral envelope of each frame is estimated by a "true" envelope of cepstral order 48 [15]. Then, the DC component of the correction filter is set to zero so as not to change frame energy. Finally, in transformation, the correction filter for the speaker is applied to the DFT magnitude spectrum of each Normal frame before synthesis.

The correction filters for each speaker are shown in Figure 4, where speakers 1-4 are male and 5-8 are female. A consistent trend that is evident is a boosting of energy in the 500-4500Hz (inclusive formant) region, confirming observations in previous works studying the Lombard effect [3]. This can be seen explicitly in the solid line in Figure 5, which shows the overall average correction filter estimated using all frames of all speakers. In Figure 4, the shape of the correction filters for male and female speakers differs somewhat, mainly due to longer/shorter vocal tract lengths of male/female speakers.

In addition to the correction filters shown in Figure 4, an "enhanced" version of each filter is also estimated for each speaker that uses only Lombard speech frames with an SII higher than 0.6. So, the average Normal spectral envelope is estimated using all frames, as this serves as a normalization for all speech frames to be transformed. Then, transformation using the enhanced filter aims to increase intelligibility as much as possible, trying to bring the transformed speech to the more focused region in which the Lombard speech is most intelligible. Such an enhancement of the correction filter essentially results in an exaggeration of the dB scale, that is, a +/-5dB gain could be scaled to +/-10dB, though the overall filter shape remains similar. The enhanced overall average correction filter is shown with the dashed-line in Figure 5. One interpretation of the correction filters, clearly observable in Figure 5, is that they serve as a sort of speaker-dependent frequency dynamic range compression, as energy is being focused more and more to a smaller range of frequencies between 500-4500Hz.

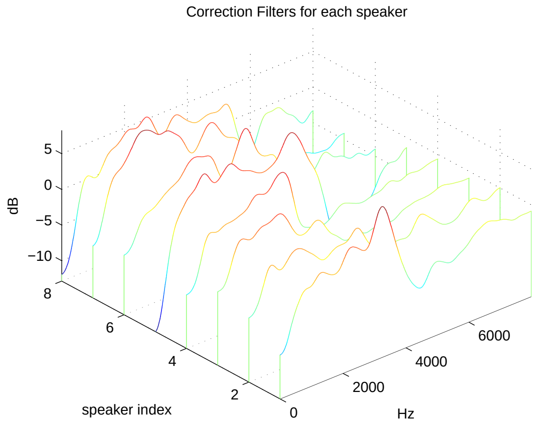

Figure 4: *Correction Filters for each speaker.*

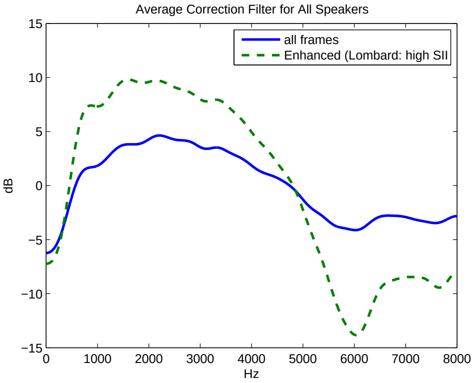

Figure 5: *Average Correction Filters, estimated from all frames (solid) and using only Lombard frames with the highest SII (enhanced - dashed line).*

### 3.1. Intelligibility of Transformed Speech

Following the analyses in previous sections, Figure 6 shows the SII histograms of the transformed speech (using both initial and enhanced versions of the correction filter for each speaker) along with the distributions for Normal and Lombard speech. The medians of the distributions are given in the Figure caption. In Figure 6, though the energy per frame of the Transformed speech is fixed to be equal to that of the Normal speech, the frames are objectively more intelligible. That is, the transformation is shown to be effective at increasing the Normal SII towards that of the Lombard speech. Moreover, the enhanced transformation is even more effective. In fact, both transfor-

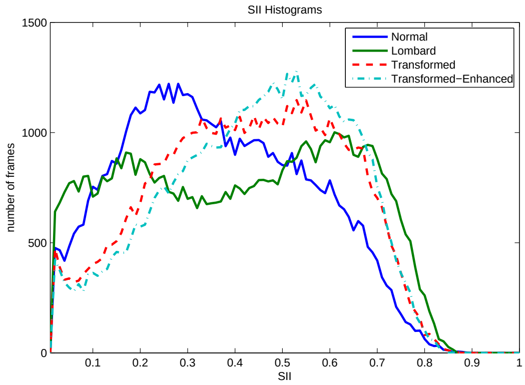

Figure 6: *SII Histograms: Normal, Lombard & Transformed speech. The medians of the SII distributions are as follows: 0.34-Normal, 0.41-Lombard, 0.42-Corrective Filters, 0.45-Enhanced Corrective Filters.*

mations yield median SIIs that are higher than that of even the Lombard speech. This can be seen in Figure 6 as a result of the transformation increasing the SII of both the voiced and unvoiced speech. Whereas it was shown in Figures 1 and 2 that the unvoiced Lombard speech is actually less loud/less intelligible than that of Normal speech.

### 3.2. Discussion

While the above evaluations objectively examine intelligibility, the question remains if these metrics correspond to subjective increases in intelligibility, as judged by human listeners. Informal listening tests on the speech samples in noise suggests that the transformed speech is indeed more intelligible than Normal speech. Further subjective evaluations are currently underway in order to confirm these observations. Additionally, considering the histograms in Figure 6, one apparent question concerns how effective or meaningful the objective SII (and Loudness in formant regions) is for evaluating unvoiced speech. For example, the intelligibility of certain unvoiced sounds, such as plosives, can depend largely on localized bursts of “noisy” energy that are not necessarily valorized by the objective Loudness and SII metrics. While these metrics are acoustically informative, their impact on subjective intelligibility merits further examination; in particular, the present work motivates careful consideration of the differences between voiced and unvoiced speech.

## 4. Conclusions & Future Work

This paper analyzed the objective loudness and intelligibility of Normal and Lombard speech in an unsupervised context, using acoustic classification based on voicing and stationarity. The main results can be summarized as follows: 1) the Lombard effect of increasing loudness/intelligibility was shown to be localized to voiced frames, with unvoiced frames actually being objectively less loud/intelligible; 2) the Lombard effect does not appear to differentiate between stationary and transient speech, though transitions are objectively less loud/intelligible for both Normal and Lombard speech. Additionally, spectral envelope transformation of the Normal speech was examined in the form of speaker-dependent correction filters that effectuated a sort of frequency dynamic range compression. The correction filter estimated using only the most intelligible Lombard speech improved the objective intelligibility most, outperforming even the Lombard speech, as both voiced and unvoiced modes demonstrated increased SII. Future work will focus on localizing acoustic characteristics of the Lombard effect to specific spectro-temporal regions in order to then develop a more precise transformation, still in an unsupervised context.

## 5. Acknowledgements

This work was supported by LISTA. The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 25623.

## 6. References

[1] E. Lombard, “Le signe de l’elevation de la voix, annals maladiers oreille,” *Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[2] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, “Effects of noise on speech producton: Acoustical and perceptual analyses,” *J. Acous. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[3] Y. Lu and M. Cooke, “Speech production modifications produced by competing talkers, babble and stationary noise,” *J. Acous. Soc. Am.*, no. 124, pp. 3261–3275, 2008.

[4] J. Junqua, “The lombard reflex and its role on human listeners and automatic speech recognizers,” *J. Acous. Soc. Am.*, vol. 93, no. 1, pp. 510–524, 1993.

[5] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Loevenbruck, “An acoustic and articulatory study of lombard speech: global effects on the utterance,” in *Interspeech*, 2006.

[6] V. Hazan and A. Simpson, “Cue-enhancement strategies for natural vcv and sentence materials presented in noise,” *Speech, Hearing and Language*, vol. 9, pp. 43–55, 1996.

[7] S. Yoo, J. Boston, A.El-Jaroudi, C. Li, J. D. Durrant, K. Kovacyk, and S. Shaiman, “Speech signal modification to increase intelligibility in noisy environments,” *J. Acous. Soc. Am.*, vol. 122, no. 2, pp. 1138–1149, 2007.

[8] B. Langner and A. Black, “Improving the understandability of speech synthesis by modeling speech in noise,” in *ICASSP*, vol. I, 2005, pp. 265–268.

[9] T. Raitio, A. Suni, M. Vainio, and P. Alku, “Analysis of HMM-based lombard speech synthesis,” in *Interspeech*, 2011, pp. 2781–2784.

[10] Y. Lu and M. Cooke, “The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise,” *SpeechComm*, no. 51, pp. 1253–1262, 2009.

[11] E. Godoy, O. Rosec, and T. Chonavel, “Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora,” *IEEE Trans Audio, Speech, Lang Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[12] D. Kapilow, Y. Stylianou, and J. Schroeter, “Detection of non-stationarity in speech signals and its application to time-scaling,” in *Eurospeech*, 1999, pp. 2307–2310.

[13] “ITU standard rec-bs.1387-1-2001,” 2001.

[14] K. Rhebergen and N. Versfeld, “A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners,” *J. Acous. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[15] A. Roebel and X. Rodet, “Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation,” in *Digital Audio Effects (DAFx)*, 2005, pp. 30–35.