Glottal-based Analysis of the Lombard Effect

Thomas Drugman, Thierry Dutoit

TCTS Lab, University of Mons, Belgium

Abstract

The Lombard effect refers to the speech changes due to the immersion of the speaker in a noisy environment. Among these changes, studies have already reported acoustic modifications mainly related to the vocal tract behaviour. In a complementary way, this paper investigates the variation of the glottal flow in Lombard speech. For this, the glottal flow is estimated by a closed-phase analysis and parametrized by a set of time and spectral features. Through a study on a database containing 25 speakers uttering in clean and noisy environments (with 4 noise types at 2 levels), it is highlighted that the glottal source is significantly modified due to the increased vocal effort. Such changes are of interest in several applications of speech processing, such as speech or speaker recognition, or speech synthesis.

Index Terms: Lombard Speech, Glottal Flow

1. Introduction

The Lombard effect, as originally highlighted by Dr. Lombard in 1909 [1], refers to the speech changes due to the immersion of the speaker in a noisy environment. In such a context, the speaker tends (generally unconsciously) to modify its way of uttering so as to maximize the intelligibility of its message [2]. On a physiological point of view, an hyper-articulation is observed when the subject speaks in noisy conditions, which is reflected by an amplification of the articulatory movements [3]. As a consequence the Lombard effect encompasses a set of acoustic and phonetic modifications in the speech signal. These modifications affect the efficiency of speech processing systems and have to be compensated for an optimal performance. Among these systems, the compensation of the Lombard effect in speech recognition [4], [5] and speaker recognition [6] have already been studied.

The analysis of Lombard speech has been studied in several works ([2], [7], [5], [3], [6]). In these studies, the acoustic and phonetic features that were inspected include the vocal intensity, phoneme durations, the fundamental frequency, the spectral tilt, and the formant frequencies. In this way, the Lombard effect is known to result in an increased vocal intensity and fundamental frequency. Duration of vowels and semi-vowels was shown to increase with the noise level, while the duration of consonants was observed to be shorter. Regarding the spectral contents, the proportion of high frequencies is more important in Lombard speech, when compared to the neutral condition [3]. This is reflected by a weaker average spectral tilt as the noise increases [2], [6]. Finally the formant frequencies were observed to be reorganized in the $F1-F2$ plane [2], [3]. While $F1$ was shown to increase in noisy conditions, no general rule were noticed for $F2$. In any of these studies, modifications were observed to be dependent on the noise type and level, as well as the considered speaker who may adapt its speaking style more or less strongly.

Among all these works, no one reported studies based on features arising from the glottal source (at the exception of the widely used fundamental frequency). However, it is expected that, during the production of speech in noise, the vocal folds work in a way different from their normal behaviour in silent conditions. Indeed, significant differences in the glottal source have already been observed between various phonation types [8]. To the best of our knowledge, only one study investigated the information extracted from the excitation for analyzing Lombard speech [9]. In that study, authors inspected the changes present in two signals: 1) a zero-frequency filtered signal, 2) the LP residual signal. Although these two signals are informative about the excitation of the vocal tract, they do not correspond to the actual glottal source produced by the vocal folds.

This paper focuses on the analysis of the Lombard speech based on features extracted from the glottal flow. This signal corresponds to the airflow arising from the trachea and modulated by the vocal folds, and is then motivated by physiological considerations. The paper is structured as follows. Section 2 describes the method we used for estimating the glottal flow directly from the speech waveform. The features that are extracted from this signal are detailed in Section 3. Section 4 presents the results of our experiments led on a large database containing an important number of speakers, noise types and levels. Finally Section 5 concludes.

2. Glottal Flow Estimation

Estimating the glottal flow directly from the speech signal is a typical problem of blind separation, since neither the vocal tract nor the glottal contributions are observable. It is then required to adopt a source-tract decomposition process incorporating a good understanding of the production mechanism. In this paper, the glottal flow is only estimated for voiced regions of speech. For such regions, Figure 1 displays the typical shape of one cycle of the glottal flow (Fig.1(a)) and its derivative (Fig.1(b)) according to the Liljencrants-Fant (LF) model [10]. Considering an abrupt return phase, the glottal cycle is composed of two main segments: the open phase and the closed phase. The closed phase refers to the timespan during which the glottis is closed. During this period, the vocal tract can be assumed to be free of any excitation. Some methods of glottal flow estimation are then based on a Closed Phase Inverse Filtering (CPIF) and aim at finding a parametric modeling of the spectral envelope computed during the estimated closed phase period [11].

In this study, the Glottal Closure and Opening Instants (GCIs and GOIs) are automatically located using the algorithm proposed in [12]. This method showed its ability to robustly determinine, in an accurate and reliable way, the GCI and GOI positions directly from the speech waveform. From them, segments of closed phase are extracted (since, considering an abrupt return phase, the closed phase is defined as the timespan between the GCI and the consecutive GOI). During the resulting closed phase, a Discrete All Pole (DAP, [13]) model is used for characterizing the vocal tract response. The DAP technique aims at computing the parameters of an autoregressive model by minimizing the Itakura-Saito distance [14], instead of the time squared error used by the traditional LPC. The Itakura-Saito distance is a spectral distortion measure arising from the human hearing perception. Finally the glottal flow is obtained by an inverse filtering step.

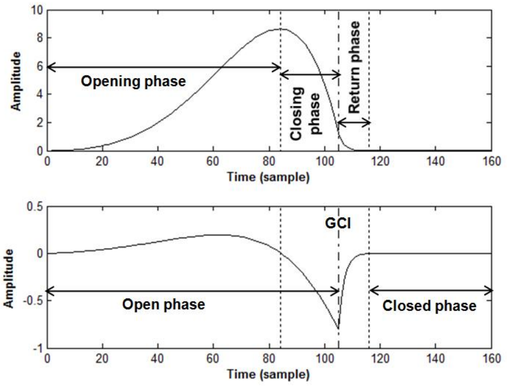

Figure 1: Typical waveforms, according to the Liljencrants-Fant (LF) model, of one cycle of: (top) the glottal flow, (bottom) the glottal flow derivative. The various phases of the glottal cycle, as well as the Glottal Closure Instant (GCI) are also indicated.

3. Characterization of the Glottal Flow

Once the glottal flow has been estimated as explained in Section 2, each glottal cycle is characterized by the following features:

- **the Normalized Amplitude Quotient (NAQ)**: NAQ is a parameter describing the glottal closing phase [15]. It is defined as the ratio between the maximum of the glottal flow and the minimum of its derivative, then normalized with respect of the fundamental frequency. Its robustness and efficiency to separate different types of phonation was shown in [15], [16]. Note that a quasi-similar feature called *basic shape parameter* was proposed by Fant in [17], where it was qualified as *"most effective single measure for describing voice qualities"*.
- **the Quasi-Open Quotient (QOQ)**: QOQ is a feature describing the relative open time of the glottis [18]. It is defined as the duration during which the glottal flow is 50% above the minimum flow, normalized to the pitch period. Note that QOQ was used in [18] for studying the physical variations of the glottal source related to the vocal expression of stress and emotion. In [16] various variants of the open quotient Oq have been tested in terms of the degree they reflect the phonation changes. QOQ was found to be the best for this task.
- **the H1-H2 ratio (H1-H2)**: This parameter is defined as the ratio between the amplitudes of the amplitude spectrum of the glottal source at the fundamental frequency and at the second harmonic [19]. It has been widely used as a measure characterizing voice quality [20], [17].
- **the Harmonic Richness Factor (HRF)**: This parameter quantifies the amount of harmonics in the amplitude spectrum of the glottal source. It is defined as the ratio between the sum of the amplitudes of harmonics, and the amplitude at the fundamental frequency [21]. It was shown to be informative about the phonation type in [21] and [22].

In this study, these features were extracted with the TKK Aparat toolkit freely available in [23]. Besides, since the fundamental frequency has been extensively used in the literature ([2], [5], [6]), the pitch information is estimated using the Snack Sound Toolkit [24]. Finally, as last feature related to the glottal behaviour, an averaged spectrum is computed for characterizing the utterances of a speaker in given recording conditions (i.e for a given noise type and level). This is achieved in a way inspired from the technique described in [6]. Voiced regions of speech are isolated and nasal segments are removed. For each resulting frame, the amplitude spectrum is computed. Periodograms are then averaged. This averaged magnitude spectrum then contains a mix of the average glottal and vocal tract contributions. If the dataset is sufficiently large and phonetically balanced, formants tend in average to cancel each other. An example of averaged spectrum for a male speaker and for three recording conditions is exhibited in Figure 2. Since these spectra were computed for the same speaker, it is reasonable to think that the main difference between them is due to the spectral tilt of the glottal flow regarding the phonation mode. In this way, it can be noticed from Figure 2 that the high-frequency contents becomes more important as the noise level increases. This confirms the conclusions about the spectral tilt drawn in [2] and [6]. In order to characterize the content of the averaged spectrum $S(f)$, the following ratios of energy are defined:

$$E_{2-1} = \frac{\int_{1000}^{3000} |S(f)|^2 df}{\int_{0}^{1000} |S(f)|^2 df}, E_{3-1} = \frac{\int_{3000}^{8000} |S(f)|^2 df}{\int_{0}^{1000} |S(f)|^2 df}.$$

The division according to these 3 subbands arises from the observation of Figure 2 where the distinction between these 3 spectral regions is well marked.

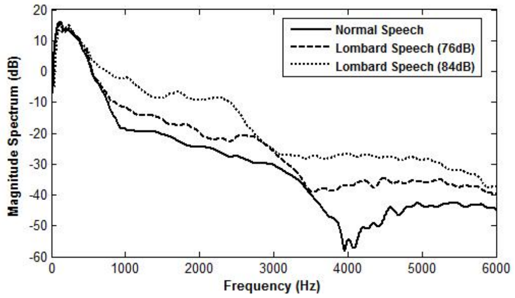

Figure 2: Averaged spectrum for a male speaker uttering in a silent environment, or with a factory noise of 76 dB and 84 dB.

4. Experiments

4.1. Database

The database used in this study was first designed by the Multitel non-profit organization in order to develop robust speech recognition systems. It consists of speech uttered by 25 speakers (11 females and 14 males). For recordings in clean conditions, the dataset consists of about 350 phonetically balanced sentences, and 57 sequences of words and numbers. For Lombard speech, four types of noise (car, crowd, factory and pop music noises) with two levels (76 and 84 dB-SPL) were used. For the noisy conditions, only the 57 sequences of words and numbers were recorded. The speech signals were captured by a close-talk microphone and sampled at 16kHz.

4.2. Results

For all recordings the glottal flow is estimated as explained in Section 2 and the features described in Section 3 are extracted from it. Among these parameters, the fundamental frequency F_0 has been extensively used in the literature. An example of F_0 distribution for a given male speaker is displayed in Figure 3 for both normal and Lombard speech. A clear increase of pitch is noticed as the noise level becomes stronger. This observation corroborates the conclusions drawn in [2], [5] or [6].

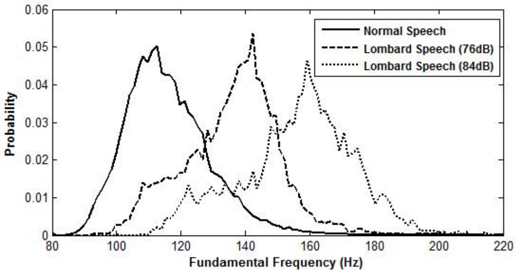

Figure 3: Pitch distribution for a given male speaker uttering in clean and noisy conditions. For this example, a factory noise at 76 and 84 dB was used.

Regarding the features characterizing the glottal waveform both in time and frequency domains, Figure 4 exhibits their histograms for the same male speaker. Also maybe less marked than for the F_0 distribution with this speaker, significant differences in the histograms of the glottal features can be nevertheless observed. In this way, the Lombard speech is characterized by a clear drop of NAQ, QOQ and H1-H2 parameters, while the Harmonic Richness Factor HRF is increased. These modifications are mainly due to the stronger vocal effort in Lombard speech, and are in line with the study of the pressed phonation type ([15], [22]).

According to the evolution of the spectral features ($H1 - H2$ and HRF), the content of the glottal spectrum is shown to present more high-frequency energy in Lombard speech. Indeed, in Lombard speech, the amplitude levels between the two first glottal harmonics becomes less important, and the amount of harmonics in the whole glottal spectrum gets richer. On the other hand, the evolution of the time-domain features NAQ and QOQ is difficult to be interpreted intuitively. To give an idea of their impact, Figure 5 displays the glottal open phase according to the LF model [10], for normal and Lombard speech, taking the mean values of NAQ and QOQ from Figure 4(a) and (b). Indeed these 2 latter parameters are known to control the shape of the glottal open phase. Differences in the glottal waveforms are observed in Figure 5, mainly in the rapidity of the open phase time response.

Table 1 sums up the modifications of glottal features when speech is produced in silent or noisy environments. These results are averaged for the 25 speakers of the database and detailed in the table according to the noise type and level. Unformally we observed that all speakers tend to modify their glottal source in the same fashion, although these changes were more important for some speakers than for others. Regarding the features extracted from the glottal flow, it turns out that the noise types leading to the strongest modifications are (by order of increasing changes): the music, crowd, car and factory noises. Besides important variations of NAQ from normal to Lombard speech can be noted (up to 26% in the factory noise at $84dB$). As expected ([2], [5], [6]), it is also observed that speakers tend to increase F_0 in Lombard speech. Finally, regarding the spectral balances E_{2-1} and E_{3-1} defined as in Equation 3, it can be concluded that speakers produce a higher amount of high-frequency in Lombard speech, confirming the results from [2], [6] and [3]. Among others, the energy in the $[1kHz - 3kHz]$ is particularly increased. One possible reason for this is that speakers (maybe unconsciously via their own auditory feedback) aim at enhancing their intelligibility by increasing the SNR where the human ear is the most sensitive.

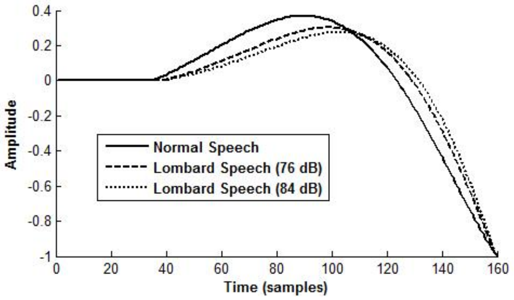

Figure 5: Illustration of the differences in the glottal open phase according to the LF model for normal and Lombard speech.

5. Conclusion

This paper investigated the modifications of the glottal source when speech is produced in noisy conditions. For this, the glottal flow was estimated by a closed phase inverse filtering process and characterized by a set of time and spectral features. Through an analysis on a database containing 25 speakers uttering in quiet and noisy environments (with 4 noise types at 2 levels), it was shown that the glottal source is considerably modified in Lombard speech. These variations have to be taken into account in applications such as speech or speaker recognition systems. Moreover the results presented in this study could be turned into advantage by integrating them in a parametric speech synthesizer based on a source-filter model. It is indeed expected that this approach should enhance the delivered intelligibility by adapting the voice quality.

6. Acknowledgments

Thomas Drugman is supported by the “Fonds National de la Recherche Scientifique” (FNRS). Authors are also thankful to Multitel ASBL for providing the Lombard database.

7. References

[1] E. Lombard, *Le signe de l’elevation de la voix*, Annales des Maladies de l’Oreille et du Larynx, vol. 37, pp 101-119, 1911.

[2] W. Van Summers, D. Pisoni, R. Bernacki, R. Pedlow, M. Stokes, *Effects of noise on speech production: acoustic and perceptual analyses*, JASA, vol. 84, pp. 917-928, 1988.

[3] M. Garnier, L. Bailly, M. Dohen, P. Welby, H. Loevenbruck H., *An Acoustic and Articulatory Study of Lombard Speech: Global Effects on the Utterance*, Proc. Interspeech, 2006.

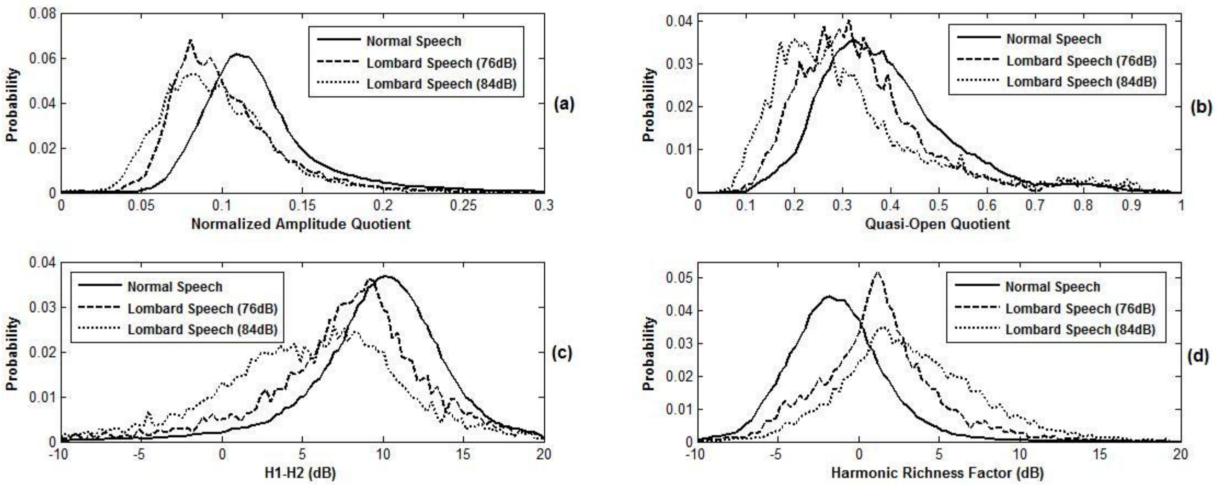

Figure 4: Distributions, for a given male speaker uttering in a quiet environment or in noisy conditions (with a factory noise at 76 and 84 dB), of the following glottal features: *(a):* the Normalized Amplitude Quotient NAQ, *(b):* the Quasi-Open Quotient QOQ, *(c):* the ratio of the amplitudes at the two first harmonics H1-H2, *(d):* the Harmonic Richness Factor HRF.

Feature	**Normal**	Car76	Car84	Crowd76	Crowd84	Factory76	Factory84	Music76	Music84
NAQ	**0.131**	-15.9%	-23.7%	-14.7%	-22.5%	-20.2%	-26.4%	-5.8%	-15.8%
QOQ	**0.411**	-7.5%	-10.5%	-4.7%	-10.8%	-9.0%	-12.6%	-2.1%	-6.6%
H1H2	**9.45 dB**	-1.8 dB	-2.3 dB	-1.8 dB	-2.5 dB	-1.9 dB	-2.9 dB	-0.6 dB	-1.1 dB
HRF	**-1.72 dB**	$+1.7$ dB	$+3.0$ dB	$+1.9$ dB	$+3.3$ dB	$+2.9$ dB	$+4.1$ dB	$+1.5$ dB	$+2.6$ dB
F0	**164.7 Hz**	$+9.8\%$	$+25.7\%$	$+20.8\%$	$+25.8\%$	$+13.7\%$	$+29.4\%$	$+25.2\%$	$+31.1\%$
E_{2-1}	**-22.62 dB**	$+8.0$ dB	$+11.5$ dB	$+8.1$ dB	$+11.0$ dB	$+9.4$ dB	$+12.8$ dB	$+10.3$ dB	$+12.6$ dB
E_{3-1}	**-28.34 dB**	$+4.6$ dB	$+6.8$ dB	$+3.7$ dB	$+4.3$ dB	$+7.8$ dB	$+10.7$ dB	$+8.3$ dB	$+9.3$ dB

Table 1: Quantitative summary of the glottal modifications in Lombard speech. The reference values of the glottal features are given for the normal speech. Their relative modifications in Lombard speech are detailed for the 4 noise types at 76 and 84 dB.

[4] H. Boril, P. Fousek, H. Hoge, *Two-stage System for Robust Neutral/Lombard Speech Recognition*, Proc. Interspeech, pp. 1074-1077, 2007.

[5] J. Hansen, *Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition*, Speech Communication, vol. 20, pp. 151-173, 1996.

[6] J. Hansen, V. Varadarajan, *Analysis and Compensation of Lombard Speech Across Noise Type and Levels With Application to In-Set/Out-of-Set Speaker Recognition*, IEEE Trans. on Audio, Speech and Language Processing, vol. 17, pp. 366-378, 2009.

[7] J. Junqua, *The Lombard Reflex and its Role on Human Listeners*, JASA, vol. 93, pp. 510-524, 1993.

[8] A. Ni Chasaide, C. Gobl, *Voice source variation*, The Handbook of Phonetic Sciences, pp. 427-461, 1997.

[9] G. Bapineedu, B. Avinash, S. Gangashetty, B. Yegnanarayana, *Analysis of Lombard Speech using Excitation Source Information*, Proc. Interspeech, 2009

[10] G. Fant, J. Liljencrants, Q. Lin, *A four parameter model of glottal flow*, STL-QPSR4, pp. 1-13, 1985.

[11] D. Wong, J. Markel, A. Gray, *Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform*, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 27, no. 4, 1979.

[12] T. Drugman, T. Dutoit, *Glottal Closure and Opening Instant Detection from Speech Signals*, Interspeech Conference, 2009.

[13] A. El Jaroudi, J. Makhoul, *Discrete All-Pole Modeling*, IEEE Trans. on Signal Processing, vol. 39, no. 2, pp. 411-423, 1991.

[14] F. Itakura, S. Saito, *A statistical method for estimation of speech spectral density and formant frequencies*, Electron. Commun. Japan, vol. 53-A, pp. 36-43, 1970..

[15] P. Alku, T. Backstrom, E. Vilkman, *Normalized amplitude quotient for parametrization of the glottal flow*, the Journal of the Acoustical Society of America, vol. 112, pp. 701-710, 2002.

[16] M. Airas, P. Alku, *Comparison of multiple voice source parameters in different phonation types*, Proc. Interspeech, pp. 1410-1413, 2007.

[17] G. Fant, *The LF-model revisited. Transformations and frequency domain analysis*, STL-QPSR, vol. 36, no. 2-3, pp. 119156, 1995.

[18] A. Laukkanen, E. Vilkman, P. Alku, H. Oksanen, *Physical variations related to stress and emotional state : a preliminary study*, Journal of Phonetics, vol. 24, pp. 313335, 1996.

[19] D. Klatt, L. Klatt, *Analysis, synthesis and perception of voice quality variations among female and male talkers*, the Journal of the Acoustical Society of America, vol. 87, pp. 820-857, 1990.

[20] H. Hanson, *Individual variations in glottal characteristics of female speakers*, Proc. ICASSP, pp. 772-775, 1995.

[21] D. Childers, C. Lee, *Vocal quality factors : analysis, synthesis, and perception*, the Journal of the Acoustical Society of America, vol. 90, no. 5, pp. 2394-2410, 1991.

[22] P. Alku, C. Magi, S. Yrttiaho, T. Backstrom, B. Story, *Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering*, the Journal of the Acoustical Society of America, vol. 125, no. 5, pp. 3289-3305, 2009.

[23] [Online], *http://aparat.sourceforge.net/index.php/Main_Page*, TKK Aparat Main Page, 2008.

[24] [Online], *http://www.speech.kth.se/snack/*, The Snack Sound Toolkit.