Linking Loudness Increases in Normal and Lombard Speech to Decreasing Vowel Formant Separation

Elizabeth Godoy[1], *Catherine Mayo*[2], *Yannis Stylianou*[1]

[1]Institute of Computer Science, Foundation of Research and Technology Hellas, Crete, Greece
[2]The Centre for Speech Technology Research, University of Edinburgh, UK

Abstract

The increased vocal effort associated with the Lombard reflex produces speech that is perceived as louder and judged to be more intelligible in noise than normal speech. Previous work illustrates that, on average, Lombard increases in loudness result from boosting spectral energy in a frequency band spanning the range of formants F1-F3, particularly for voiced speech. Observing additionally that increases in loudness across spoken sentences are spectro-temporally localized, the goal of this work is to further isolate these regions of maximal loudness by linking them to specific formant trends, explicitly considering here the vowel formant separation. For both normal and Lombard speech, this work illustrates that, as loudness increases in frequency bands containing formants (e.g. F1-F2 or F2-F3), the observed separation between formant frequencies decreases. From a production standpoint, these results seem to highlight a physiological trait associated with how humans increase the loudness of their speech, namely moving vocal tract resonances closer together. Particularly, for Lombard speech, this phenomena is exaggerated: that is, the Lombard speech is louder and formants in corresponding spectro-temporal regions are even closer together.

Index Terms: Lombard Effect, Loudness, Vowel Formant Separation

1. Introduction

The Lombard effect describes how humans reflexively modify their speech when speaking in a noisy environment [1]. The increased vocal effort associated with the Lombard reflex produces speech that is perceived as louder and that is more intelligible to listeners when presented in noise. Many works have studied the acoustic-phonetic modifications associated with the Lombard effect [2, 3, 4, 5, 6] and increased vocal effort [7]. In particular, the Lombard decrease in spectral "tilt" has been shown to be essential in capturing the perceived tenseness of the style. This spectral trend increases loudness and augments the speech intelligibility in noise [5, 8]. Moreover, the observed decrease in spectral tilt and corresponding increase in loudness can further be linked acoustically to characteristics of the glottal flow [9, 10]. In addition to the average spectra, previous studied have also examined trends in the average formant values for Lombard speech compared to its normal counterpart [3, 4, 11]. Findings from these works indicate a clear increase in F1 for Lombard speech, often associated with the increase in fundamental frequency, as well as a decrease in F3, with the results for F2 being mixed. Ultimately, the majority of studies on Lombard speech focus on average acoustic-phonetic trends. However, important acoustic-phonetic cues in speech are rapidly varying across sentences, within words, syllables and even phones. Accordingly, a primary motivation underlying this work is to localize these observed spectral and formant trends further, examining isolated spectro-temporal regions of maximal loudness within vowel segments.

In particular, previous work in [8] highlighted the Lombard increase in loudness for voiced speech via boosting spectral energy, on average, in an inclusive (roughly 500-4500Hz) frequency band, effectively making formants more audible. Localizing these analyses further, Fig. 1 provides an example of Lombard and normal speech from [8], with the calculated loudness shown along with the spectral envelopes across the sentence (cf [8] and Section 2 for more details), focusing on an inclusive formant band. Note that the rms-energies of the sentences are equal. The main point to be gleaned from Fig. 1 is that the regions of maximal loudness, for both the normal and Lombard speech, are spectro-temporally localized. Moreover, the loudness patterns are similar for both sentences, with the loudness of the Lombard speech being noticeably greater (or more pronounced). Examining Fig. 1 more closely, it is clear that the loudness does not increase uniformly across the speech, but is rather isolated to lower or upper frequency bands containing F1-F2 or F2-F3, respectively. More specifically, there appears to be a relationship between these formants, specifically their movement and separation, and the increases in loudness. Following these observations, this work seeks to explicitly isolate these regions of maximal loudness observed in normal and Lombard speech and link them to formant trends, focusing particularly here on vowel formant separation. From a speech production standpoint, the hypothesis underlying this work is that loudness increases as formants move closer together, indicating that humans are physiologically adjusting their vocal tracts to this end. This increase in loudness can then be linked to decreased spectral tilt, potentially stemming from the glottal source (as suggested in other works, though not examined explicitly here).

In order to begin illustrating this hypothesized phenomena, the present work takes several steps, outlined as follows. First, Section 2 describes the normal and Lombard speech corpora and details on calculating features used in analyses. Section 3 begins by explaining how speech frames with maximal loudness (in an inclusive formant band) are localized within vowel segments and shows the loudness distributions for these frames compared to all of the normal and Lombard speech (from vowels). The corresponding average spectra and vowel spaces are then examined, showing decreased spectral tilt and a clear shifting of formants with increasing loudness. Next, Section 4 explicitly examines the vowel formant separation and then further localizes the regions of maximal loudness to F1-F2 and F2-F3 bands. Ultimately, in using these bands and also in isolating

Thanks to EU Future and Emerging Technology (FETOPEN) Project LISTA (The Listening Talker) for support and encouragement.

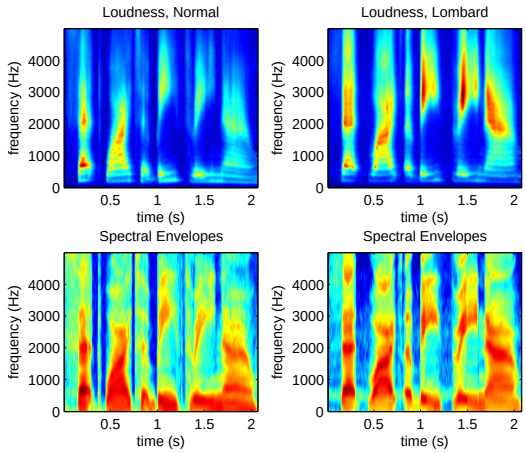

Figure 1: A motivating example illustrating the calculated loudness (above) and spectral envelopes (below) for the sentence "set white at B three now," spoken by a female in a normal (left) and Lombard (right) voice.

only the maximal loudness from the overall band distributions, it is progressively illustrated that, as loudness increases, the corresponding vowel formant separation decreases. Finally, Section 5 concludes and discusses implications for future work.

2. Speech Corpus and Processing

2.1. Grid Lombard Corpus

The Lombard (and normal) speech data is from the Grid corpora presented in [12, 4, 5]. The sentences have a simple 6-word structure (e.g. "place red in G 9 soon"), as defined in the Grid multitalker speech corpus [12]. Each sentence was read and recorded both in quiet conditions (normal) and while the speaker listened through headphones to speech-shaped noise at a 96dB level (Lombard). In this work, the Lombard speech corresponding to the highest noise level (i.e., Ninf96) in [4, 5] was selected so that the Lombard reflex characteristics would be most obvious. Finally, 50 sentences per speaker, from 8 British English speakers (4 male, 4 female) are considered in analyses and the speech sampling rate is 16kHz, downsampled from 25kHz.

2.2. Speech Signal Analyses and Segmentation

The speech signal analysis is pitch-asynchronous (using a 30ms Hanning window and a 10ms step) and DFT-based with an FFT length of 2048, while the spectral envelope of each frame is estimated by a "true" envelope of cepstral order 48 [13]. Moreover, each Lombard sentences is rms-normalized to match its respective normal counterpart. In this way, the energy differences between the conditions are normalized across the sentences. Also, it should be noted explicitly that, unlike the analyses in [8], the Lombard speech in this work is not time-aligned to the normal speech. Lastly, for the speech segmentation, and automatic HTK-based audio-to-text aligner, provided by the University College London (UCL) Speech, Hearing & Phonetic Sciences Department, was used and no manual corrections were performed.

2.3. Loudness and Formant Estimation

Loudness is a psychoacoustic descriptor of a signal's impact on the human auditory system. While several works propose models to capture and quantify loudness of audio signals and speech, there is no absolute metric. However, the International Telecommunication Union (ITU) standards for loudness calculation seeks to lay out relevant criteria and common guidelines. The Perceptual Evaluation of Audio Quality (PEAQ) metric follows the pertinent ITU standard [14] and is described in detail in [15]. The basic version of PEAQ is used in this work (and in [8]), which employs an FFT-based ear model and considers the signal energy and modulations in critical auditory bands after applying outer and middle ear frequency responses. In particular, as in [8], the Loudness considered here is the spline-interpolated PEAQ metric, averaged across a frequency range spanning F1-F3 (i.e., 500-4000Hz). Later in Section 3, this average Loudness will be split into two bands representing F1-F2 and F2-F3 ranges, respectively.

In addition to Loudness, this work focuses on formant trends, specifically considering vowel spaces and vowel formant separation. Accordingly, the formants are estimated frame-by-frame via a Least Squares Auto-Regressive fit (AR order 16) of the spectral envelope. Specifically, the formants are estimated in a basic manner as the AR poles with frequency greater than 90Hz and bandwidth less than 400Hz. Then, in analyses, formant estimates corresponding to either F1 greater than 1000Hz or F2 less than 800Hz are not considered. Additionally, in all formant-related statistics presented in this work, a trimmed mean is employed (keeping 90% of the given feature data, i.e. removing 5% of the samples from the upper and lower extremities of the data distribution) in order to limit the influence of potential outliers.

3. Illustration of Increasing Loudness and Corresponding Vowel Space Shifts

3.1. Isolating Maximal Loudness in Vowel Segments

The analysis approach adopted in this work is original, in that analyses are localized to within phone segments, specifically isolating a frame for each vowel with maximal Loudness. As mentioned in the previous section, to begin with, the Loudness represents an average over a "full" 500-4000Hz inclusive F1-F3 formant band. The main novelty in this work then lies in comparing the "peak" Loudness frames with "all" of the segment frames, for both normal and Lombard speech, so that spectral and formant trends corresponding to increasing loudness can be highlighted, as is shown explicitly next.

3.2. Loudness Distributions and Average Spectra

First, Fig. 2 shows the loudness distributions for the normal-N and Lombard-L speech, comparing the distributions calculated using "all" vowel segment frames to those calculated using only the "peak" Loudness frames (i.e, the single frame per vowel segment with maximal loudness). The Loudness distributions are histograms that are normalized by the total number of frames considered (so that each distribution sums to one). As can be seen from Fig. 2 a clear progression of increasing loudness is established, beginning with Normal-all, then Normal-peak,followed by Lombard-all and Lombard-peak.

Additionally, Fig 3 shows the overall average spectral envelope calculated using the frames from each of the distributions in Fig 2. As can be seen unequivocally from Fig 3, the progression of increasing loudness involves a decrease in spectral "tilt," or more specifically, an increase in spectral energy mainly between 1-4.5kHz (with a decrease in energy near DC or 0Hz). Related work in [9, 10] would further suggest that

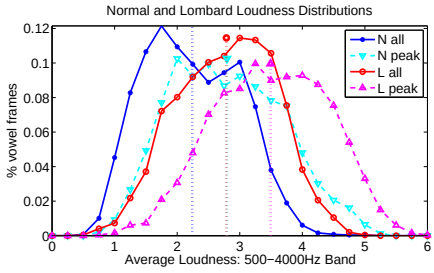

Figure 2: Loudness Distributions in Full Formant Band (500-4000Hz) for normal-N and Lombard-L speech, with "peak" indicating frames with maximal Loudness versus "all" vowel segment frames. Distribution means are shown by vertical lines with height matching the maximum.

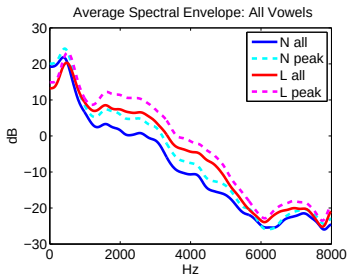

Figure 3: Average Vowel Spectral Envelopes (corresponding to the distributions shown in Fig. 2).

this observed trend in average spectra with increasing loudness originates from the glottal source spectral tilt.

3.3. Vowel Space Movement with Increasing Loudness

Similarly to the illustration of decreasing spectral tilt with increasing loudness, the vowel space "movement" or shifts with increasing loudness are shown in Fig 4. Each vowel space point is calculated as a trimmed mean (90% of the formant values are kept). The vowel space area is represented using the convex hull (i.e., a polygon fit that encompasses all of the data points) in order to capture the maximal area that the points span. Additionally, while the vowel space typically refers to only the F1-F2 plane, this work also examines the F2-F3 plane. In the end, to the authors' knowledge, this type of vowel space analyses focusing on increasing loudness is unique to the present work.

As can be seen from Fig 4, there is a clear shifting of vowel spaces with increasing loudness. In particular, F1 is increased and F3 is decreased, with F2 moving up or down, depending on the vowel. These observations are in accordance with the average formant values found in studies of Lombard speech. Here, however, there is a clear view of the entire F1-F2 and F2-F3 planes, respectively observing a stretching out of the F1-F2 hull tip to the right and movement downwards of the spaces in the F2-F3 plane with increasing loudness. Moreover, considering the all-to-peak movement observed for both normal and Lombard speech, the formant shifts are even more pronounced (than going from normal-to-Lombard speech considering all vowel segment frames). What is even more interesting is that the influence of fundamental frequency is limited in this former case, as the same vowel segments are considered. That is, while the F1 increase in Lombard speech is often attributed to an increase in fundamental frequency, the all-to-peak progression here shows that F1 increases with increasing loudness, even when considering the same vowel segments within a given style of speech.

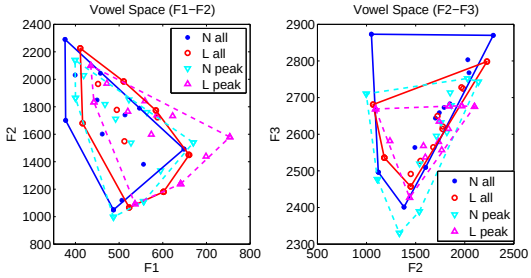

Figure 4: Vowel Spaces in the F1-F2 and F2-F3 planes for the same conditions examined in Fig 2-3.

4. Linking Increased Loudness to Decreased Vowel Formant Separation

Observing the vowel space movement with increasing loudness illustrated in the previous Section, the hypothesis developed that could explain these trends is that loudness peaks or maximizes when formants move closer together. Specifically, when F1 is close to F2 or when F2 is close to F3. The analyses in this Section illustrate trends that support this hypothesis.

4.1. Average Formant Separation

First, Table 1 shows the average vowel formant separation corresponding respectively to the frames considered in the distributions from the previous Section: that is, normal-all, normal-peak, Lombard-all and Lombard-peak. For each vowel, the average formant separation is a trimmed mean (90% kept) of the calculated formant distances. The values in Table 1 then represent the average formant separation across the vowels. As can be seen clearly in comparing Fig 2 and Table 1, with increasing loudness, formant separation decreases (i.e. formants are closer together overall).

Table 1: Average Formant Separation in Hz. Note that the peak Loudness considers the full (500-4000Hz) band. Relevant comparisons with indicated conditions are provided in parentheses (italics) in order to clearly quantify certain differences.

	F2-F1	**F3-F2**
N all (Na)	1186	962
L all (La)	1115 (-71 Na)	943 (-8 Na)
N peak (Np)	1132 (-54 Na)	938 (-24 Na)
L peak (Lp)	1089 (-26 La)	928 (-15 La)

4.2. Isolating Loudness in F1-F2 and F2-F3 Formant Bands

Furthermore, considering the example shown in Fig 1 and the hypothesis made at the beginning of this Section, the following analyses consider the Loudness split into broad "lower" and "upper" formant bands respectively capturing F1-F2 and F2-F3. That is, the analyses conducted previously are repeated, but with the Loudness now covering 300-2000Hz for F1-F2 and 1500-3500Hz for F2-F3. In this way, the Loudness increases are further localized in frequency. For example, considering the diphthong in Fig 1 beginning at about 0.5 sec, a peak in Loudness in the lower F1-F2 band will be found at the beginning of the segment while a peak in Loudness in the upper F2-F3 band will be found at the end of the segment.

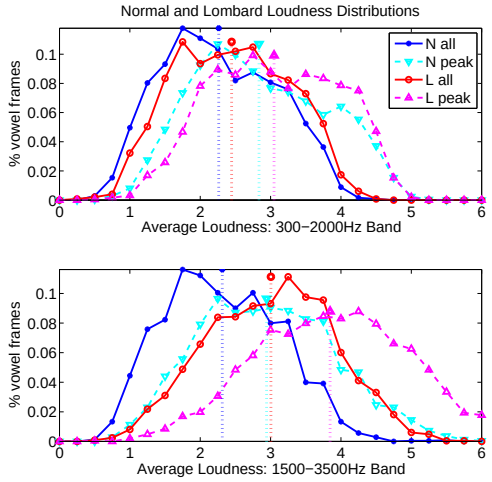

Figure 5: Distributions of Average Loudness calculated over the Lower F1-F2 (above) and Upper F2-F3 (below) Formant Bands.

Similarly to Fig 2, Fig 5 shows the Loudness distributions, where Loudness is calculated for the F1-F2 band (above) and F2-F3 band (below). The same progression of increasing Loudness as Fig 2 can be seen in Fig 5. One notable observation is that the difference between the peak Loudness distributions, compared to all, are more pronounced in the F2-F3 band, with the peak Loudness being greater for both styles of speech.

Like Fig 4 and 2, Fig 6 displays the vowel spaces corresponding to the Loudness distributions shown in Fig 5, considering the F1-F2 (above) and F2-F3 (below) bands. The "peakL" and "peakU" labels emphasize the fact that the peak Loudness respectively considers the Lower and Upper bands in these cases. What can be seen clearly now is that the increase in Loudness (progression from N-all to L-peak) for the F1-F2 band corresponds to an increase in F1 and decrease in F2. On the other hand, considering the vowel spaces with increasing Loudness in the F2-F3 band, there is a significant decrease in F3, with F2 tending to increase. These observations seem to support the hypothesis that Loudness peaks in spectro-temporal regions for which F1 and F2 or F2 and F3 are close together. Like Table 1, Table 2 provides the average formant separation (in Hz) for the peak Loudness in the F1-F2 (Lower-L) and F2-F3 (Upper-U) bands. Along with the observations from Fig 6, the values in Table 2 confirm that Loudness peaks in the lower/upper bands are directly linked to the separation between F1-F2/F2-F3 decreasing.

Table 2: Average Formant Separation in Hz, peaks in Loudness averaged over Lower(L)/Upper(U) Bands.

	F2-F1	F3-F2
N peakL	**1068** (-118 Na)	961
L peakL	**1005** (-110 La)	971
N peakU	1189	**895** (-67 Na)
L peakU	1121	**888** (-55 La)

4.3. Formant Separation for Maximal Loudness in Bands

Finally, in order to further emphasize the observed trends supporting the hypothesis that Loudness maximizes when either F1 and F2 or F2 and F3 are close together, Table 3 shows the average formant separation (similarly to Table 2) though considering only the maximally Loud frames in the Lower and Upper bands. That is, instead of using all of the frames from the Loudness distributions in Fig 5, only the frames above the distribution mean are used in calculating the formant separation values shown in Table 3. As can be seen from Table 3, as the "Loudest" regions in the F1-F2 and F2-F3 bands correspond to F1-F2 and F2-F3 being the closest together, respectively.

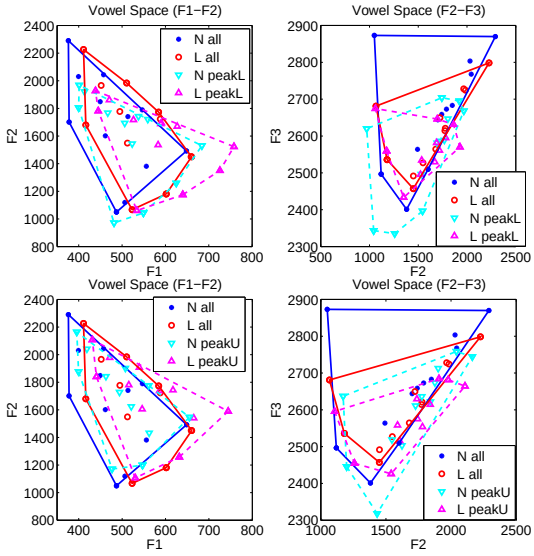

Figure 6: Vowel Spaces, peak Loudness averaged over the Lower F1-F2 (above) and Upper F2-F3 (below) bands.

Table 3: Average Formant Separation in Hz, considering Maximal-M Loudness from peaks in Lower(L)/Upper(U) Band Distributions.

	F2-F1	F3-F2
N MpeakL	**1052** (-134 Na)	979
L MpeakL	**962** (-153 La)	968
N MpeakU	1259	**829** (-133 Na)
L MpeakU	1208	**808** (-135 La)

5. Conclusions and Discussion

This work illustrates an apparent physiological mechanism of human speech production, namely that loudness peaks in spectro-temporal regions where formants (either F1-F2 or F2-F3) are close together. Moreover, the closer the formants, the louder the region. In the case of Lombard speech, this phenomena is exaggerated.

From a speech modification perspective, the above observations could be used to tailor spectral tilt modifications to the formant information (e.g. decreasing spectral tilt further when formants are found to be close together). In this way mimicking a physiological trait, the speech modifications might increase both naturalness and intelligibility.

In addition to the examination of vowel formant separation in this work, future work will consider the formant bandwidth and also formant dynamics in order to examine if the increases in loudness are indeed linked to the formant movement, in addition to their separation. Additionally, analysis of formant dynamics (e.g. vowel formant transitions) would bring a significant feature of speech perception into the forefront, effectively examining if increases in loudness are linked to emphasizing important perceptual cues.

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] 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.

[4] 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.

[5] ——, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *SpeechComm*, no. 51, pp. 1253–1262, 2009.

[6] B. J. Stanton, L. H. Jamieson, and G. D. Allen, "Acoustic-phonetic analysis of loud and lombard speech in simulated cockpit conditions," in *ICASSP*, 1988, pp. 331–334.

[7] J. S. Lienard and M. G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *J. Acous. Soc. Am.*, no. 106(1), pp. 411–422, 1999.

[8] E. Godoy and Y. Stylianou, "Unsupervised acoustic analyses of normal and lombard speech, with spectral envelope transformation to improve intelligibility," *Interspeech, Portland Oregon, USA*, 2012.

[9] T. Drugman and T. Dutoit, "Glottal-based analysis of the lombard effect," in *Interspeech, Japan*, 2010.

[10] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *J. Acous. Soc. Am.*, vol. 126, no. 4, pp. 2061–2071, 2009.

[11] C. Davis and J. Kim, "Is speech produced in noise more distinct and/or consistent?" in *Speech Science and Technology*, 2012, pp. 46–49.

[12] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition (l)," *J. Acoust. Soc. Am., Letters to the Editor*, vol. 120, no. 5, pp. 2421–2424, 2006.

[13] 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.

[14] "ITU standard rec-bs.1387-1-2001," 2001.

[15] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, and et. al., "PEAQ–the ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29, 2000.