To appear in: Proc. Interspeech 2014, Singapore

Acoustic characteristics of critical message utterances in noise applied to speech intelligibility enhancement

Neehar Jathar, Preeti Rao

Department of Electrical Engineering, Indian Institute of Technology Bombay, India

neehar.j@iitb.ac.in, prao@ee.iitb.ac.in

Abstract

Motivated by the potential of speech modification for the enhancement of intelligibility in noisy environments, we study the acoustic characteristics of speech produced in the context of critical announcements made in noisy listening situations. A corpus of 3 speakers producing 20 Marathi train station announcements is analysed for articulatory-acoustic and prosodic differences between speech in noise and in quiet. It is observed that apart from the global changes that are characteristic of increased vocal effort, the intonation associated with phrase and word accents is consistently modified. Listening tests with modified speech suggest that spectral shaping and F0 modifications that are linguistically valid contribute constructively to increased intelligibility in noise, as measured in an information extraction task.

Index Terms: speech intelligibility enhancement, Lombard effect, prosody

1. Introduction

Speech communication can be severely compromised in noisy listening situations, as is well known. There has been active interest in the speech processing community to enhance speech signals to increase robustness to anticipated degradations due to background noise [1][2][3]. High intelligibility is the single most important requirement for speech reaching a listener in adverse environments such as broadcast announcements in a public space. It has thus been of great interest to examine how humans adapt their speech production to such challenging communication requirements. The Lombard reflex describes how humans involuntarily modify their speech in a noisy environment by increasing vocal effort in order to aid communication. Lombard speech has been demonstrated to provide improved intelligibility in noise at the same SNR as normal speech [4]. Several studies have examined the acoustic characteristics of Lombard speech in various languages. Changes have been noted not only in terms of signal amplitude in dB SPL, but also in formant frequencies and bandwidths, spectral tilt, durations (overall and distribution across vowels and consonants) and fundamental frequency (F0) [5][6][7][8]. Some of the observations have been quantified and applied to transform speech to simulate the corresponding aspects of the Lombard effect.

Cooke et al. [9] investigate various speech modifications using time-frequency energy reallocation to obtain intelligibility improvements with natural and synthetic speech. Godoy et al. [2] exploit the observed consistency in relative energy increase in the formants band in Lombard speech to derive a spectral shaping filter and demonstrate gains in intelligibility at low SNRs. Jokinen et al. [3] apply spectral tilt compensation, formant sharpening and high-frequency emphasis via an adaptive post-filter to improve intelligibility of telephone speech.

While modifications to the speech spectrum to increase intelligibility have been studied extensively, the prosodic aspects of Lombard speech have been relatively less researched. Increase in mean F0 with respect to normal speech is a commonly noted aspect but modified speech with global F0 scaling has been found to not affect intelligibility [10]. In fact, F0 modulation results by Miller et al. [11] concluded that non-linguistic modifications reduce intelligibility. Motivated by linguistic aspects, Garnier et al. [12] showed that prosodic cues to phrasing and word segmentation are enhanced in Lombard speech. Davis et al. [13] modelled the F0 contours of normal and Lombard English sentences using the Fujisaki model to find an increase in phrase and accent command amplitudes. Since it is known that F0 provides important linguistic cues to speech perception, it is of interest to determine how F0 modifications can be effectively used to improve speech intelligibility in noise.

This work seeks to quantify the change in the F0 contour from normal to Lombard speech in terms of Fujisaki model parameters [14]. This parameterization helps provide the continuous F0 contours required for the synthesis of Lombard speech from normal utterances in quiet. We have also investigated the feasibility of a speaker-dependent filter (but fixed across phones) that can be applied to normal speech to achieve the spectral envelope of the corresponding Lombard speech. The above analysis has been carried out with Marathi, an Indian language with close to 70 million native speakers, in the context of a communication task involving critical information about train arrivals at a station. Such messages are commonly conveyed by an announcer in a quiet location over a public address system and received by the target audience immersed in high background noise. Suitable speech modification before broadcast over the public address system has the potential to increase its communication effectiveness.

The organization of the paper is as follows. Section 2 describes the dataset and discusses listening tests conducted to confirm the intelligibility gain of Marathi Lombard speech. The modelling of the prosodic as well as the spectral changes in Lombard speech with reference to normal speech is presented in Section 3. The results of the listening test with Lombard-inspired modified speech are summarized in Section 4 while Section 5 presents the conclusions.

2. Dataset and perceptual testing

Normal and Lombard speech was recorded and listening tests were carried out to verify the intelligibility benefit of Lombard speech for Marathi.

2.1. Corpus

The corpus was a set of 10 Marathi sentences. The template, borrowed from actual announcements routinely heard at train stations in Mumbai city, was “Train number <a><b><c><d> down, <name> express, will halt at platform number <y>

instead of platform number <x>". The words <a>, <b>, <c> and <d> consisted of a digit from 0-9, the train name was chosen from among five choices (all 3 syllables long) and <x>, <y> consisted of a number between 10-19. The utterances present a high cognitive load for the listener and thus seem ideal for use in intelligibility testing.

Table 1: *List of information words in the corpus*

Train numbers	Train names	Platform numbers
Shunya (0)	Rajdhani	Dahaa (10)
Ek (1)	Gorakhpur	Akraa (11)
Don (2)	Duronto	Baaraa (12)
Teen (3)	Chandigarh	Teraa (13)
Chaar (4)	Mangalore	Choudaa (14)
Paach (5)		Pandhraa (15)
Sahaa (6)		Solaa (16)
Saat (7)		Satraa (17)
Aath (8)		Athraa (18)
Nau (9)		Ekonees (19)

An example sentence in Marathi - *"Train kramaanka* ***paach ek saat don*** *down,* ***Rajdhani*** *express, platform kramaanka* ***solaa*** *chya aivaji, platform kramaanka* ***chouda*** *var thambel."*

The sentences were designed such that each train number digit occurred in each of the four word positions and each platform number occurred in each of the two positions. This was vital from the point of view of the listening tests of Section 2.3.

2.2. Audio recording

A total of 3 native Marathi speakers, 2 male and 1 female were recorded. Brown noise was played out through SONY MDR-XD200 headphones to the speakers at two levels, 30 dB apart. The lower level of noise was barely perceptible and corresponded to the normal speech condition. The speakers were given no specific instructions regarding intelligibility. The 10 sentences were spoken by each speaker at the two noise levels producing normal and Lombard utterances. These were recorded in a quiet room using a SHURE SM58 microphone at 44.1 kHz and later down-sampled to 16 kHz for the analyses.

Brown noise was used as the stimulus for eliciting the Lombard effect because of its stationary nature and spectral similarity with actual train station noise recorded during peak hours using a ROLAND R-09HR handheld recorder. The long term average spectrum of the recorded train station noise is shown in Figure 1.

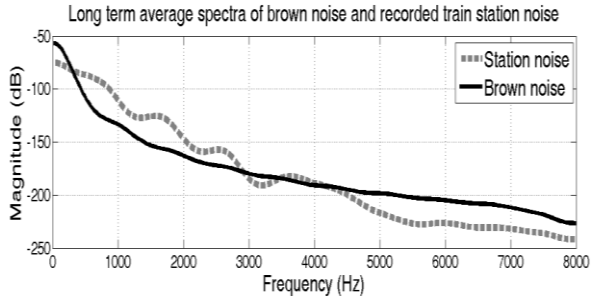

Figure 1: *Long term average spectra of brown noise and recorded train station noise*

2.3. Listening Tests

2.3.1. Design

Utterances for listening tests had the same template as the recorded sentences but were synthesized by inserting randomized train numbers, names and platform numbers extracted from the set of full utterances of one of the male speakers. To make sure that the word position related prosodic effects were undisturbed, we ensured that a particular slot in the template was filled by a digit/word uttered in the same slot in the recording. A relatively stationary sample of noise from the train station recording was added to the synthesized utterances at two levels of SNR, -9 and -6 dB. Thus normal and Lombard utterances in the listening test had the identical overall SNR. These SNRs were chosen to create conditions similar to the speaker's environment in Section 2.2.

Each participating listener heard 20 sentences – 10 normal and 10 Lombard, equally distributed over the 2 SNRs. The listener had to identify the 4 digits of the train number, the train name and the two platform numbers and key in the responses before moving on to the next sentence.

2.3.2. Results

A total of 18 people participated in the listening tests, either native Marathi speakers or residents of Mumbai (where the native language is Marathi) for over 10 years. Word error rates (WER) were computed separately for train numbers, train names and platform numbers. As expected, train name WER was lowest as compared to train numbers and platform numbers because the names were 3 syllables long and chosen from among only 5 options. Platform number WER was high due to similar sounding words such as *akraa* (11) and *athraa* (18) which differ in only 1 consonant.

The improvement in WER for Lombard speech over normal speech was evident in all cases. The relative decrease in WER from normal to Lombard was similar across SNRs for each of the categories as seen in Table 2.

Table 2: *WER (%) from listening test with 18 listeners*

	SNR	Low Vocal Effort	High Vocal Effort
Train names	-6 dB	5.9	2.4
	-9 dB	28.2	12.9
Train numbers	-6 dB	25.3	13.2
	-9 dB	51.2	26.8
Platform numbers	-6 dB	38.2	24.7
	-9 dB	75.3	45.9

3. Acoustic Analyses

The changes in acoustic characteristics of Lombard speech with respect to normal speech were studied by observing pairs of identical utterances by each speaker under the two conditions. Overall mean intensity was higher in the Lombard case than in the normal case for all speakers. Unlike most other researchers, we found that the utterance duration decreases marginally from the normal to the Lombard case. These results are consistent with the results presented in [15] who have also used full sentence speech data and attributed the decrease to a "sense of urgency" expressed by the speaker. The mean fundamental frequency in the Lombard case was higher than in the normal case, but there was high inter-

speaker variability in the extent. The overall prosodic analysis results appear in Table 3.

Table 3: *Changes in prosodic parameters from normal case to Lombard case for three speakers*

Speaker	Mean Phrase Duration	Mean F0	Mean Intensity
1 (Male)	-8%	+31%	+8 dB
2 (Male)	-2%	+54%	+9 dB
3 (Female)	-2%	+7%	+2 dB

3.1. F0 analysis using Fujisaki modelling

With a view to applying the analysis outcomes to speech modification where a continuous F0 contour must be specified, the Fujisaki model [14] was chosen as the underlying representation. Equally importantly, the model is interpreted in terms of phrase and accent commands, which automatically constrains modifications of model parameters to be linguistically valid. In the Fujisaki model, the pitch contour is represented as a superposition of a phrase component and an accent component on a base frequency. The phrase and accent components are represented as second order linear filters, excited by an impulse and a rectangular function respectively. The amplitude of the accent excitation is denoted by A_a while the amplitude of phrase excitation is denoted by A_p. The base frequency is denoted by F_b. An example of F0 contour representation using the Fujisaki model is shown in Figure 2. The applicability of the Fujisaki model to Marathi has not been studied earlier although there is past work on modelling prosody of another Indian language, Hindi [16].

Phone-level segmentation was carried out on each utterance using a Marathi speech recognition system to obtain the corresponding syllables in the sentence pair. The base frequency, F_b, of the normal utterance was used as the F_b of the Lombard utterance as well. The number of phrase commands and accent commands were fixed for a given pair of utterances. If an accent command or phrase command was missing, a dummy command with zero amplitude was inserted so that direct comparisons could be made between the normal and Lombard case. As observed from the ratios presented in Table 4, the phrase command amplitudes A_p values increase significantly from the normal to the Lombard case and in a manner related inversely to the mean A_p in the normal speech of the corresponding speaker. This could be a consequence of the average normal pitch being on the higher side of the total pitch range for certain speakers leaving smaller room for increase in the Lombard case. The analysis of accent commands, unlike phrase commands, requires two parameters, amplitude and duration. The dummy accent commands with zero amplitude were not included while calculating the averages. Table 4 summarizes the observations of all the Fujisaki parameters.

Table 4: *Fujisaki parameter summary for normal and Lombard speech*

Speaker	F_b (Hz)	Mean normal A_p	Mean A_p ratios	Mean normal A_a	Mean A_a ratios	Mean duration ratio for accents
1	90	0.37	1.76	0.30	1.23	1.17
2	80	0.27	2.79	0.35	1.4	1.25
3	160	0.60	1.33	0.55	1.05	1.13

While there is not much literature on lexical stress in Marathi, it is not considered distinctive. However, stressed syllables do appear in focus words following linguistic rules [17] and were observed in our data as such. We find that while all speakers show increase in mean A_a and mean duration of the accent commands, the increase is not as large as in the case of phrase commands. On observing the word accents in the Lombard speech, it was noted that these are not restricted to the information words. Further, the relative change in duration of accent commands tends to vary widely across syllables due to the occurrence of reduced vowels that compensate for stressed vowels in their vicinity. Hence, the change in duration of accent commands was not incorporated in the synthesis of Lombard speech presented in Section 4.

3.2. Spectral analysis

It is known that the Lombard effect causes a redistribution of spectral energy from lower to higher frequencies. A spectral shaping filter to convert the average normal voice spectrum to the Lombard voice spectrum was calculated separately for each speaker, using a method similar to Godoy et al. [2].

First, the frame-wise spectral envelope was obtained by using the first 40 cepstral coefficients for 20 ms Hamming windowed data with a 10 ms hop across the utterance. It was next averaged over the complete recorded data for a given speaker at a given vocal effort. This was done for both the normal and Lombard utterances giving two long term average energy spectra. The bin-wise ratios of the Lombard spectrum magnitude to the normal spectrum magnitude gave the frequency response of the required corrective spectral shaping filter. The filters for the three speakers are shown in Figure 3. We note a distinct increase in spectral energy in the 2-4 kHz region for all speakers. Both the male speakers' filters show a peak between 3-4 kHz, while for the female speaker the peak is shifted rightwards to between 4-5 kHz. For all speakers, energy at almost all frequencies above 1 kHz increases in the Lombard case, consistent with the reduced spectral tilt due to increased vocal effort. However, the magnitude of the Lombard effect in terms of the overall spectral energy redistribution varies considerably with the speaker.

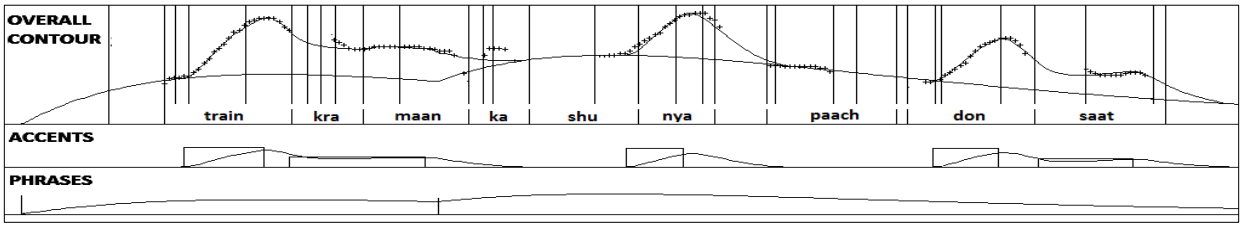

Figure 2: *Pitch contour representation using the Fujisaki model. The lowest pane shows the phrase component, the one above that shows the accent component while the topmost pane shows the sum of the two along with the original pitches (+)*

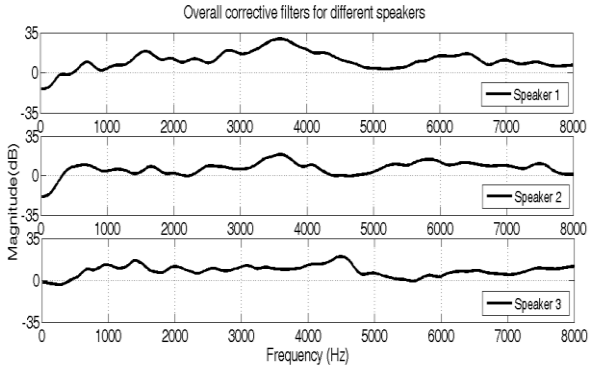

Figure 3: *Overall corrective filters for the three speakers*

Next, the similarity of the spectral shaping filter across individual vowels was observed. The long term average energy spectra were calculated in the same manner as before but for four separate groups of vowels (front high, front low, mid and back) separately, using the available annotated data. Figure 4 shows the vowel-wise corrective filter in comparison with the overall corrective filter for Speaker 1.

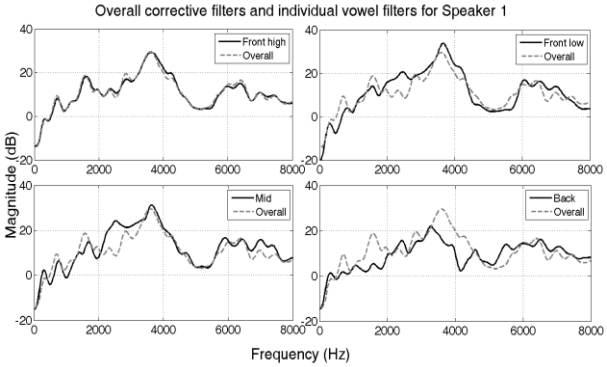

Figure 4: *Average corrective filter shape along with vowel specific corrective filters for Speaker 1*

We see that the overall corrective filter is similar to the vowel specific filters for all the vowel groups, with an increase in energy in the 2-4 kHz region and an overall reduction in spectral tilt, justifying the use of a single filter. The average filter matches the front high vowels more closely than it does the back vowels because the data is dominated by the former.

4. Intelligibility of modified speech

The normal utterances of Speaker 1 were modified based on the acoustic analyses of the previous section in order to study the effect on intelligibility. The A_p values were scaled uniformly by a factor of 1.78 and the A_a values by a factor of 1.23 leaving accent durations unchanged. Utterances were re-synthesized using the new pitch contour thus obtained, using the TD-PSOLA algorithm [18]. Further, these utterances were passed through the fixed overall corrective filter calculated for Speaker 1 in Section 3.2. A listening test was conducted along the lines of the original listening test using four types of utterances, viz. original normal, normal with F0 modification, normal with spectral shaping and normal with both spectral and F0 modifications. As before, the listening test SNR was uniform across all normal and modified utterances.

The intelligibility testing was restricted to train numbers (due to the higher reliability of scores) and the results appear in Table 5. The WER for the unmodified normal speech was similar to the results of the original listening test. We observed a large decrease in WER from the normal speech to the fully modified case. Only-F0-shifted utterances showed no significant change in WER as compared to the normal utterances. Spectral shaping alone, however, affected WER favourably. The combined spectral and F0 modification showed a significant WER improvement over normal speech in the -6 dB SNR case, but not so in the -9 dB SNR case. We speculate that the increased F0 presents a trade-off between improved prosodic cues causing an increase in intelligibility and reduced harmonic density in the spectrum causing a decrease. While these results are preliminary, they indicate a potential benefit of F0 modifications to intelligibility.

Table 5: *WER (%) from listening test with 10 listeners for train numbers only*

SNR	Normal speech	F0 modified	Spectrum modified	F0+spectrum modified
-6 dB	24.5	24	19.5	13
-9 dB	56.5	54	38.5	39.5

5. Conclusion

In this work, an analysis of Marathi Lombard speech with reference to normal speech was carried out to determine the acoustic correlates of the observed improved intelligibility of Lombard speech. Analysis of the prosody confirmed the rise in mean pitch and intensity that has been widely reported, but Lombard utterances were found to be marginally reduced in duration. The fundamental frequency contour was represented using the Fujisaki model and the average phrase command amplitudes as well as the accent command amplitudes were observed to increase for Lombard speech. The average spectral energy redistribution varied across speakers but an increase in spectral energy in the formant regions between 2-4 kHz was found to be common. Finally, a listening test was conducted in realistic background noise with synthesized speech based on modelling the intonation and spectral changes observed in natural Lombard speech. Spectral modifications resulted in a drastic improvement in intelligibility while prosodic changes caused a smaller but definite improvement at selected SNR. The results demonstrate that linguistically valid modifications to F0 can enhance intelligibility improvements normally obtained by spectral energy redistribution alone. Future work would focus on replicating the speech modification results on a larger set of speech and noise data. The synthesis of F0 modifications, which was observed to be slightly unnatural due to artefacts characteristic of TD-PSOLA, needs improvement especially for large F0 scaling factors [19]. Finally, other aspects of intelligibility as observed in clear speech articulation could be included in speech modification for intelligibility enhancement [2].

6. Acknowledgements

We would like to thank Dr. H. Mixdorff for his advice regarding the Fujisaki model and the Bharti Centre for Communication at IIT Bombay, who supported this work in part.

7. References

[1] M. Cooke, C. Mayo and C. Valentini-Botinhao, “Intelligibility enhancing speech modifications: the Hurricane Challenge”, in *Proc. Interspeech*, Lyon, France, 2013.

[2] E. Godoy, M. Koutsogiannaki and Y. Stylianou “Approaching speech intelligibility enhancement with inspiration from Lombard and Clear speaking styles”, *Computer Speech & Language*, vol. 28, no. 2, pp 629-647, 2014.

[3] E. Jokinen, M. Takanen, M. Vainio and P. Alku, "An adaptive post-filtering method producing an artificial Lombard-like effect for intelligibility enhancement of narrowband telephone speech.", *Computer Speech & Language*, vol. 28, no. 2, pp 619-628, 2014.

[4] A. L. Pittman, and T. L. Wiley, "Recognition of speech produced in noise." *Journal of Speech Language and Hearing Research*, vol. 44, no. 3, pp 487-496, 2001.

[5] D. Pisoni, R. Bernacki, H. Nusbaum and M. Yuchtman, “Some acoustic – phonetic correlates of speech produced in noise”, *Int. Conf. on Acoust. Speech and Sig Proc*, 1985, pp. 1581 – 1584.

[6] J. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers”, *J. Acoust. Soc. Am.*, vol. 93, pp. 510–524, 1993.

[7] 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 *Proc. Interspeech*, 2006.

[8] L. Youyi and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise." *J. Acoust. Soc. Am.*, vol. 124, 3261, 2008.

[9] Y. Tang, and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions." In *Proc. Interspeech*, 2010.

[10] L. Youyi and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise." *Speech Communication*, vol. 51, no. 12, pp 1253-1262, 2009.

[11] S. E. Miller, R. S. Schlauch and P. J. Watson, "The effects of fundamental frequency contour manipulations on speech intelligibility in background noise", *J. Acoust. Soc. Am*, vol. 128, no. 1, 2010.

[12] M. Garnier, M. Dohen, H. Loevenbruck, P. Welby and L. Bailly, "The Lombard Effect: a physiological reflex or a controlled intelligibility enhancement?", Proc. *7th ISSP* , 2006, pp 255-262.

[13] C. Davis, J. Kim, K. Grauwinkel and H. Mixdorff, "Lombard speech: Auditory (A), Visual (V) and AV effects.", *Proc 3rd Int. Conf. on Speech Prosody*, 2006.

[14] H. Fujisaki, S. Ohno and S. Narusawa, “Physiological mechanisms and biomechanical modelling of fundamental frequency control for the common Japanese and the standard Chinese,” in *Proc. 5th Seminar on Speech Production*, Bavaria, Germany, 2000, pp 145-148.

[15] J. Hansen and 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. Audio, Speech, and Lang. Proc.*, vol. 17, no. 2, pp 366-378, 2009.

[16] H. Fujisaki and S. Ohno, “Analysis and modeling of fundamental frequency contours of Hindi utterances”, in *Proc. Interspeech*, 2005, pp 1413-1416.

[17] R. V. Dhongde and K. Wali, *Marathi*, Philadelphia, PA: John Benjamins Publishing Co., 2009, 2, 5, pp 18-20.

[18] F. Charpentier and E. Moulines, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”, *Speech Communication*, Vol. 9, no. 5, pp 453-467, 1990.

[19] E. Moulines and J. Laroche, “Non-parametric techniques for pitch-scale and time-scale modification of speech”, *Speech Communication*, Vol. 16, no. 2, pp 175-205, 1995.