Contribution of $F_0$ Contour Level, $F_0$ Contour Shape and Durations Towards Perception of Lombard Speech

Botsa Kishore Kumar$^1$, Paidi Gangamohan$^2$, and Suryakanth V Gangashetty$^3$

$^1$International Institute of Information Technology, Hyderabad, India
$^2$Koneru Lakshamaiah Education Foundation, Hyderabad, India
$^3$Koneru Lakshamaiah Education Foundation, Vaddeswaram, India

kishore.botsa@research.iiit.ac.in, ganga39@klh.edu.in, svg@kluniversity.in

Abstract

Lombard speech, that is speech spoken in noise, has considerable changes in the speech production mechanism both at source level and system level compared to speech produced in normal/neutral speech. At source level, previous studies have shown that there is increase in fundamental frequency ($F_0$), but it is unclear that the importance of $F_0$ contour shape and $F_0$ contour level in lombard speech. These perceptual studies carried out in this paper highlight the role of $F_0$ contour shape and $F_0$ contour level along with the durations. A set of six different experiments were conducted by modifying durations and both $F_0$ contour shape and level. These experiments are performed on samples from the Hurricane Challenge-2013 database. Perceptual results show that the contribution of $F_0$ contour level is more than $F_0$ contour shape and durations.

Index Terms: Lombard speech, fundamental frequency, duration, speech synthesis

1. Introduction

For better communication in noisy environments, the humans change their speaking style [1], called lombard speech [2]. It is uncertain what characteristics of lombard speech really contribute to improvement in intelligibility. Further, the relation between these production characteristics and the noise is unknown. Resolving this problem by modifying speech allows better communication in many practical scenarios like railway stations, traffic signals, etc.

Lombard speech has been analyzed for extraction of speech production features that are making it different from normal [3]. A few studies indicate that there is an increase in intensity or loudness of lombard speech. Along with increase in intensity, considerable changes in the spectrum are observed [4]. It is observed that there is an increase in spectral flatness, due to changes in the pattern of vocal fold vibrations [5]. At phonetic level [6], the energy shifts to vowels from consonants and energy of semi-vowel shifted to consonants and vowels. Studies [7,8] on formants of lombard speech have shown that there is an increase in F1. A few studies report that there are changes in harmonics. In [8], it was observed that there is decrease in H1-H2 ratio (ratio between energy at $F_0$ and energy of second harmonic), and there is increase in spectral flatness in the glottal spectrum.

Studies [9, 10] investigated lombard speech durations with normal speech at word level and syllable level. It was shown that there is increase in the durations in either case. But study [8] showed that vowel durations got reduced and consonant durations were prolonged. This leads to a decrease in speaking rate.

Although, studies [9, 11, 12] reported that there is an increase in $F_0$ but there is no clear evidence that it contributes to intelligibility. The current study investigates $F_0$ contour shape and $F_0$ contour level towards perception of lombardness. Along with $F_0$ contour shape and $F_0$ contour level, durations are also analyzed. In this study, lombard speech is modified using neutral speech $F_0$ contour shape, level and durations. From the perceptual analysis on synthesized signals, it is observed that $F_0$ contour level carries lombard characteristics.

This paper is organized as follows: Section 2 describes the data set used in the study. In Section 3, methods for modification of $F_0$ contour shape and $F_0$ contour level and durations are given. Section 4
Figure 1: DTW Alignment of lombard utterance and neutral utterance. (a) Neutral speech signal, (b) lombard speech signal, (c) alignment path, and (d) mapped pitch contour.

gives the summary of the study.

2. Database

In this study, the Hurricane Challenge-2013 [13] database is used. It consists of 2542 plain sentences and 720 lombard sentences. The data is collected from a single male British speaker and lombard speech is recorded while the speaker was listening to speech-shaped noise (SSN) at 84dB(A). For this study, 12 sentences each from plain speech and lombard speech are used. That is 12X2 = 24 utterances are used.

3. Modification of Durations and $F_0$

The analysis-by-synthesis experiments are performed using a flexible analysis-by-synthesis tool (FAST tool) [14]. The following are the four components modified using the FAST tool:

- Dynamic vocal tract system
- Excitation source information
- Duration information
- Intonation

Speech production mechanism is characterized by the dynamic vocal tract system and the excitation source. Durations and intonation will add the suprasegmental characteristics that are acquired by an individual over a period of time.

In the FAST tool, the dynamic vocal tract information is extracted from the linear prediction (LP) analysis. It is represented by a sequence of frame LP coefficients (LPCs) obtained by using a frame size of 20 msec and a frame shift of 3 msec. The residual obtained from the LP analysis is used as an excitation source.

The $F_0$ contour of a speech signal is obtained by using the zero frequency filtering (ZFF) method [15]. The ZFF method is used to extract the epoch locations (the locations where the vocal tract is excited by impulse-like excitation). This is done by passing the speech signal through a cascade of two ideal zero frequency resonators followed by trend removal. Trend removal includes local mean subtraction, measured at all sampling instants over a period of time which is about 1.5 times the mean pitch period of the signal. The output of trend removal operation is referred to as the ZFF signal. The epoch
locations are identified as negative to positive zero crossings in the ZFF signal and referred to as glottal closure instants. The duration between the two consecutive epochs is referred to as instantaneous pitch period \( T_0 \), and \( F_0 = 1/T_0 \).

The source and target speech signals are aligned by using the dynamic time warping (DTW) algorithm. The weighted linear prediction cepstral coefficients (wLPCCs) of source and target signals are used for the alignment. The extent of duration modification is given by the alignment path between the source and target. The alignment of source (lombard speech) and target (neutral speech) signals are shown in Fig. 1. The procedure [16] for modification of components is as follows:

Step-1: Excitation source information in the form of LP residual and vocal tract system information in the form of LPCs are given by LP analysis.

Step-2: Modify epoch sequence of source according to the target prosody (duration and \( F_0 \) contour level and shape).

Step-3: Determine the modified LP residual signal from the transformed epoch sequence.

Step-4: Use the transformed LP residual and the LPCs to generate the new speech signal.

3.1. Experimental Set up

This study is carried out by a set of 6 different experiments. The experiments are:

**E1**: Flattening the lombard speech \( F_0 \) contour using averaged \( F_0 \).

**E2**: Modifying the lombard speech \( F_0 \) contour shape and level as that of neutral speech \( F_0 \) contour shape and level.

**E3**: Modifying the lombard speech \( F_0 \) contour shape as that of neutral speech \( F_0 \) contour shape by retaining the \( F_0 \) contour level of lombard speech.

**E4**: Modifying the lombard speech durations as that of neutral speech durations using alignment path as shown in Fig 1(c).

**E5**: Modifying the lombard speech \( F_0 \) contour shape and \( F_0 \) contour level as that of neutral speech \( F_0 \) contour shape and level along with durations.

**E6**: Modifying the lombard speech \( F_0 \) contour shape and durations as that of neutral speech \( F_0 \) contour and duration by retaining lombard speech \( F_0 \) contour level

3.2. Perceptual Evaluation

Perceptual evaluations of modified speech signals are done by 16 volunteers (generally, called subjects) which includes speech researchers and students. These subjects are asked to listen to original lombard speech and modified speech, and give similarity in lombardness scores on the scale of 1 to 5. A score of 5 represents the lombardness in the modified speech signal is almost similar to that of original lombard signal and score 1 represents the lower lombardness in the modified speech signal.

3.3. Discussions

The perceptual scores are listed in Table 1, where column A represents scores of modified speech signal and columns B, C, D represent perceptual scores of modified speech signal added with -5dB, 0dB and 5dB noise levels. E1-E6 correspond to experiments. The scores of experiment E1 i.e., making \( F_0 \) contour flat, are very low compared to all other modifications. This gives an idea that there is still some information of \( F_0 \) contour that is contributing to the lombardness.

The perceptual scores of experiment E2 are on the lower side, indicating that modifying both \( F_0 \) contour shape and level as that of neutral speech gives less lombardness in the speech. In experiment E3, the modified speech signals are close to lombard speech, where \( F_0 \) contour shape is modified as that of neutral speech by retaining \( F_0 \) contour level of lombard speech. This indicates that the \( F_0 \) level contributes more towards perception of lombard speech. Experiment E4 which is only modifying lombard speech durations as that of neutral speech, gives the similarity closer to lombard speech. Perceptual scores for E5 and E2 are similar (towards the lower side) this is because of the \( F_0 \) contour level is modified in both. Perceptual scores of experiment E6 are slightly on the higher side although the \( F_0 \) contour shape and durations are modified but the \( F_0 \) contour level is retained.

From E1, it is observed that making \( F_0 \) flat will degrade lombardness and it is an evidence that \( F_0 \) contributes to lombardness in some way. From experiments E3, E4 and E6, where the perception of
Table 1: Perceptual scores of modified speech signal, and at different noise condition (A: Modified speech signal, B: Modified speech signal added with -5dB noise, C: Modified speech signal added with 0dB noise, D: Modified speech signal added with 5dB noise)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment-1 (E1)</td>
<td>2.16</td>
<td>1.33</td>
<td>1.92</td>
<td>2.04</td>
</tr>
<tr>
<td>Experiment-2 (E2)</td>
<td>3.2</td>
<td>1.58</td>
<td>2.67</td>
<td>3.28</td>
</tr>
<tr>
<td>Experiment-3 (E3)</td>
<td>4.16</td>
<td>1.92</td>
<td>3.5</td>
<td>4.04</td>
</tr>
<tr>
<td>Experiment-4 (E4)</td>
<td>4.1</td>
<td>1.87</td>
<td>3.37</td>
<td>3.91</td>
</tr>
<tr>
<td>Experiment-5 (E5)</td>
<td>3.25</td>
<td>0.88</td>
<td>2.73</td>
<td>3.33</td>
</tr>
<tr>
<td>Experiment-6 (E6)</td>
<td>3.83</td>
<td>1.67</td>
<td>3.13</td>
<td>3.5</td>
</tr>
</tbody>
</table>

A Lombard in speech is high, where $F_0$ contour level is retained with $F_0$ contour shape of neutral.

The study further included the combinational analysis of durations and both $F_0$ contour level and shape. It is observed that the modified Lombard speech with the neural speech durations carried Lombardness as original Lombard speech and suggests that the duration has less effect on Lombard level. It is also observed that among $F_0$ contour level, $F_0$ contour shape and durations, $F_0$ contour level seems to be contributing to maintaining the Lombardness.

4. Summary and Conclusion

In this paper, contribution of $F_0$ contour level, $F_0$ contour shape and durations towards perception of Lombard speech is carried out. Experiments have shown that the synthesized Lombard speech signal with average $F_0$ contour (flat $F_0$ contour) doesn’t carry the Lombard information. But retaining the $F_0$ contour level with neural speech $F_0$ contour shape gives the better Lombard level. Further, duration modification is done independently and in composition with $F_0$ contour. From results it is observed that durations have very little effect on the Lombard level.

This study can be extended to analyze the effect of other source and spectral characteristics. Further, compositional analysis with different characteristics can be done.

5. Acknowledgements

The authors would like to thank the Department of Science and Technology (DST), India for supporting Paidi Gangamohan through the project SRG/2020/001363.

6. References


