INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speech Synthesis in Noisy Environment by Enhancing Strength of Excitation and Formant Prominence

Bidisha Sharma and S. R. Mahadeva Prasanna

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati-781039
{s.bidisha,prasanna}@iitg.ernet.in

Abstract

Text-to-speech (TTS) synthesis systems have grown popularity due to their diverse practical usability. While most of the technologies developed aims to meet requirements in laboratory environment, the practical appliance is not limited to a specific environment. This work aims towards improving intelligibility of synthesized speech to make it deployable in realism. Based on the comparison of Lombard speech and speech produced in quiet, strength of excitation is found to play a crucial role in making speech intelligible in noisy situation. A novel method for enhancement of strength of excitation is proposed which makes the synthesized speech more intelligible in practical scenario. Linear-prediction analysis based formant enhancement method is also employed to further improve the intelligibility. The proposed enhancement framework is applied in synthesized speech and evaluated in presence of different types and levels of noise. Subjective evaluation results show that, the proposed method makes the synthesized speech applicable in practical noisy environment..

Index Terms: Text-to-speech synthesis, intelligibility, enhancement, strength of excitation, formants

1. Introduction

Research in TTS is progressing to make synthesized speech more like natural speech. To replace human speaker by a TTS system in practical environment, it must have flexibility to adapt various manipulation to synthesized speech based on practical scenario. Human beings are flexible to change their speech signal characteristics in practical situation like noisy environment, by changing articulatory movement for the ease of the listener's perception. There are two possibilities to achieve this with a TTS: first is to record hyper-articulated speech and develop TTS using that. However, recording of hyper-articulated database may be a complex process and the level of articulation to be controlled may be based on user environment. The second way is to modify existing TTS synthesized speech to make it more intelligible in noisy environment. In noisy scenario, Lombard speech is produced by hyper-articulation to make it intelligible by compensating the audible disturbances introduced by noise [1]. The extent of hyper-articulation depends on level of noise present in the environment. To make TTS system deployable in a practical noisy scenario, there is requirement of adapting characteristics of Lombard speech to synthesized speech [2]. Lombard speech is more intelligible compared to speech produced in quiet [3, 4]. Various studies are done in the literature to compare different attributes of Lombard speech to that of normal speech, some of which can be enhanced to make the speech signal sound like Lombard speech [5]. Lombard speech is found to have more duration, pitch and less spectral tilt compared to speech produced in quiet. The level of these modifications depends on the extent of noise present in the environment [6–8]. Methods like consonant-vowel (CV) energy ratio boosting, spectral shapers, high-pass-filtering followed by amplitude compression are found to enhance intelligibility to significant extent [9–12]. There are several works done in application of *speech synthesis in noise*, which also aims to modify spectral and temporal attributes to enhance the intelligibility of synthesized speech. Significant improvement in intelligibility is achieved in [13], in presence of speech shaped noise by modifying Mel-cepstral-coefficients. Spectral shaping techniques with energy re-allocation from higher to lower frequency is found to improve intelligibility in stationary and speech shaped noise [14]. Further, [15] makes an effort to add change in duration, fundamental frequency and spectral tilt to increase intelligibility of synthesized speech. A highly intelligible hidden Markov model (HMM) based speech synthesizer is developed by adapting the Lombard speech and also by modifying the vocoder attributes [16]. Along with modification of vocoder parameters, various other modifications like duration, pitch, spectral tilt, harmonic-to-noise ratio, formant enhancement are also implemented and reported to improve intelligibility compared to natural speech in low SNR condition. A useful database for studying speech synthesis in noise, (CMU_SIN) is described in [17], where first 500 utterances of CMU ARCTIC dataset have been recorded with and without noise. For noisy condition, low level babble noise is played through headphone to the voice talent during recording. In [18], speaking rate and fundamental frequency are analyzed for CMU_SIN database and temporal modification like modifying dynamic range of speech signals are performed.

In all the works described above, different modifications like duration, pitch, different characteristics of vocal-tract-spectrum are extensively performed to achieve adequate intelligibility of synthetic speech in presence of noise. As of now, no modification of source parameters are found to be reported except the fundamental frequency (F_0). As in Lombard speech, due to hyper-articulation the glottal closure becomes sharper, there are significant changes in source attributes also [19]. Authors in [20] have done some useful analysis of excitation source characteristics in Lombard speech. Duration, pitch, strength of excitation and loudness parameter are compared across Lombard and normal speech. Significant difference in the distribution of strength of excitation and loudness are observed in [20]. However, loudness modification or strength of excitation enhancement is not attempted in the field of speech synthesis in noise. Analysis and enhancement of source characteristics may be an interesting and useful cue for enhancing intelligibil-

http://dx.doi.org/10.21437/Interspeech.2016-1005

ity or adapting source characteristics of Lombard speech. In this work, source characteristics are analyzed and compared across Lombard speech and speech produced in quiet for CMU_SIN database. Based on the observations, a novel source enhancement method is proposed to modify synthesized speech and make it more intelligible in noisy environment. A linear prediction (LP) based formant enhancement is also employed to further boost the enhancement.

The rest of the paper is arranged in the following sections: Section 2 compares characteristics of source in Lombard speech and speech produced in quiet. Based on the observations, source modification of synthesized speech is performed in Section 3. Section 4 describes formant prominence enhancement. Experimental evaluation is explained in Section 5. Section 6 discusses the conclusion and future direction.

2. Analysis of strength of excitation for Lombard speech

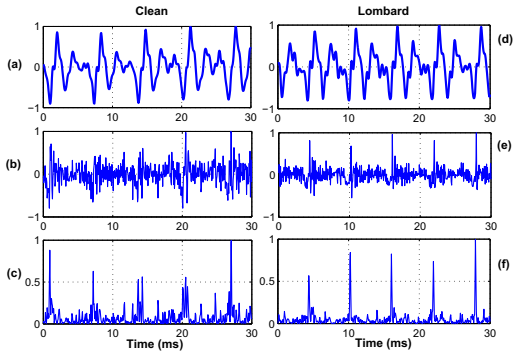

Figure 1: *(a) 30 ms speech segment (b) LP residual (c) HE of LP residual for speech produced in quiet; (d) 30 ms speech segment (e) LP residual (f) HE of LP residual for speech produced in noise*

The main focus of this work is to enhance source aspects and formant prominence of synthesized speech to make it intelligible in noisy situation. In this regard, CMU-SIN database is used for analysis [17], as it contains 500 sentences of the same speaker in noisy and quiet environments. It is convenient for comparison due to the same speaker's same utterances in both conditions. Moreover, the database is specifically designed for speech synthesis in noise. Since this database is recorded in presence of very low level babble noise, the entire database cannot be termed as Lombard speech. However, by manual listening it is found that some speech files are very loud and Lombard effect is prominent in those cases. By listening to the entire database and comparing utterances in noisy and clean environment, 200 utterances are selected out of 500 in each condition which are found to have significant difference during perception and have effect of hyper-articulation. In this study, the manually selected utterances produced in noise are termed as Lombard speech and corresponding utterances produced in quiet are termed as normal speech.

LP residual is a useful approximation of time varying excitation source of speech signal, where sharp discontinuities can be observed at glottal closure instants (GCIs), either in positive or negative polarity. This behavior of impulse-like excitation can be better visualized and quantified from Hilbert envelope (HE) of LP residual of speech signal. Sharpness of these peaks also correlates to loudness parameter as described in [21]. HE ($h_e(n)$) of LP residual ($e(n)$) is defined as follows:

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{1}$$

where $e_h(n)$ is Hilbert transform of $e(n)$ and in given by

$$e_h(n) = IDFT[E_h(k)] \tag{2}$$

where,

$$E_h[k] = \begin{cases} -jE(k) & k = 0, 1, ...(\frac{N}{2}) - 1; \\ jE(k) & k = (\frac{N}{2}), (\frac{N}{2}) + 1, ...(N-1) \end{cases} \tag{3}$$

Here, IDFT denotes inverse discrete Fourier transform and $E(k)$ is computed as discrete Fourier transform (DFT) of $e(n)$ and N is number of points for computing DFT.

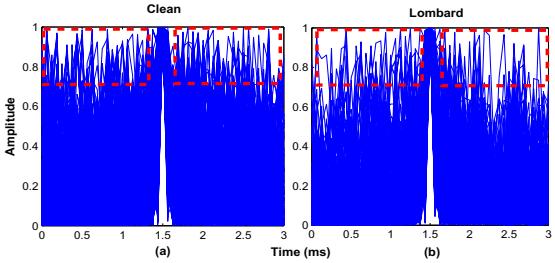

Figure 2: *Superimposed segments of Hilbert envelope of LP residual in the vicinity of impulse-like excitations for (a) speech produced in quiet, (b) Lombard speech.*

Figure 1(a) and (d) depict 30 ms speech segments corresponding to speech produced in clean and in noisy environments respectively, which correspond to same sound unit from first utterance of CMU_SIN database. Figure 1(b) and (e) show corresponding LP residuals and (c), (f) shows HE of LP residuals for normal and Lombard speech respectively. The sharp discontinuities in LP residual of Lombard speech segment in Figure 1(e) are more emphasized in HE of LP residual, as shown in Figure 1(f). The peaks of HE of LP residual in case of Lombard speech are more close to impulse-like excitations, which has maximum strength. The spread and distribution of energy around GCIs seem to be less in case of Lombard speech. The spread of energy around GCIs can be a representation of strength of excitation of a speech signal [21]. For better interpretation of this fact, 3 ms segments of HE of LP residual around each GCI are superimposed over each other. Each segment is normalized with respect to maximum value of the segment. The resultant plot is shown in Figure 2 where (a) shows the superimposed plot for all voiced frames of speech produced in quiet and same for Lombard speech is shown in (b). It is evident from dotted regions of Figure 2, that the energy associated with the side lobes around main lobe (corresponds to GCI) of normal speech is more compared to that of Lombard speech. To quantify this, entire 3 ms of each segment is divided into three equal parts; average energy associated with each region be E_1, E_2 and E_3. Then, ratio of energy of middle 1 ms segment, to sum of energy of both side segments (each of 1 ms) is calculated as $\beta = \frac{E_2}{E_1+E_3}$. The distributions for this β corresponding to Lombard and normal utterances are shown in Figure 3, which shows significant difference between both the classes for voiced segments. However, in case of unvoiced segments, no clear distinction in β is observed. Hence, for later part of this work excitation strength is modified only in case of voiced regions of synthesized utterances.

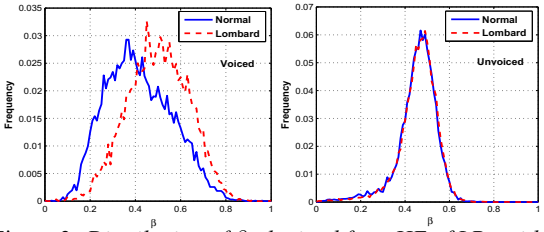

Figure 3: *Distribution of β obtained from HE of LP residual for voiced and unvoiced frames in case of Lombard and normal speech*

3. Enhancement of strength of Excitation

As per the evidences obtained in Section 2, sharpness of peaks in HE of LP residual is an important cue of hyper-articulation. In is section, effort towards modifying excitation source or strength of excitation of synthesized speech to make it similar to excitation source of Lombard speech is explained in detail. LP residual represents excitation source and its samples are highly uncorrelated, therefore robust to modification to some extent. The first step towards enhancement of LP residual would be locating GCIs accurately, which is performed by employing zero frequency filtered signal (ZFSS), where positive zero crossings of ZFFS can be represented as GCIs [22]. For introducing a large discontinuity at each GCI, residual signal around the GCI is multiplied by a Gaussian window function (w) with mean μ and variance σ. Let us consider, number of GCIs in the given utterance be N. To derive residual signal $r(n)$, the speech signal is passed through LP inverse filter. The samples of $r(n)$ ($r(n_i)$) around i^{th} GCI is enhanced as follows:

$$r_{en}(n_i) = \quad r(n_i) * w \tag{4}$$

where $n_i = (e_i - \frac{l}{2}), ...e_i, ...(e_i - \frac{l}{2})$, e_i are epoch locations, $i = 1, 2, ...N$ and l is length of the Gaussian shaped window. Minimum and maximum values of the window can be selected depending on the required amount of enhancement. Similarly, variance can also be changed. In this work, $l = 0.5\ ms$, $\sigma = 1$ and amplitude of the window varies from 1 to 3. These parameters are selected experimentally in such a way, that the energy of the side-lobes gets de-emphasized and energy associated with the main-lobe (at each GCI) gets more emphasized. The enhanced residual is passed through the previously obtained LP filter to derive the enhanced speech signal. The enhanced speech seems to be more intelligible in presence of noise. This is evident from the Figure 4, where the LP residual in Figure 4(e) clearly have sharper discontinuities compared to that of Figure 4(b). Again, same is visible from Figure 4(c) and (f). In Figure 4, (a), (b), (c) corresponds to 30 ms speech segment, its LP residual and HE of LP residual respectively for normal speech, while (d), (e), (f) depicts same for enhanced speech. In Figure 5(a) and (b), the narrow-band spectrogram corresponding to LP residual of a normal utterance and enhanced residual is shown respectively. The harmonics of source in voiced portions are darker in the spectrogram corresponding to enhanced LP residual compared to that of normal speech.

The above discussion establishes the enhancement of strength of excitation of speech produced in quiet to make it similar to speech produced in noise. After achieving the first level of enhancement with respect source, the enhanced speech is further subjected to formant enhancement to increase intelligibility.

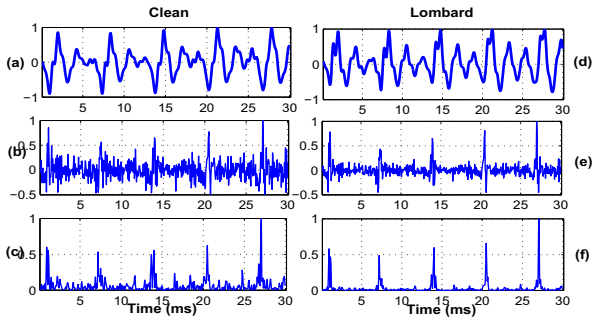

Figure 4: *(a) 30 ms speech segment (b) LP residual (c) HE of LP residual for speech produced in quiet; (d) 30 ms speech segment (e) LP residual (f) HE of LP residual for enhanced speech using proposed method*

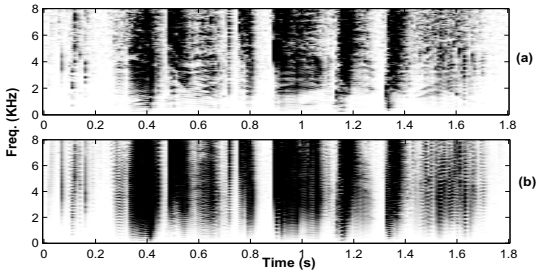

Figure 5: *Narrow-band spectrogram for LP residual of (a) Speech produced in quiet, (b) Enhanced speech.*

4. Enhancement of formant prominence

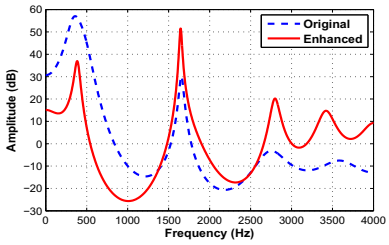

Figure 6: *Original and enhanced log-magnitude LP spectrum for 20ms speech segment*

Formant prominence also plays a vital role in perceived intelligibility of speech. Moreover, concentration of energy over spectral range, where human auditory system is most sensitive, also improves intelligibility. Based on these two facts, enhancement of formants based on LP analysis is followed in this work [23]. Firstly, speech signal is pre-emphasized and fed to LP inverse filter obtained from first order LP analysis. As pre-emphasis increases energy at higher frequency and first order LP analysis models the spectral tilt, the residual signal obtained by passing speech signal through first order LP inverse filter will have more higher frequency components. Further, using this residual, $(\frac{f_s}{1000} + 4)$ order LP analysis is performed to model LP spectrum with formant peaks and more higher frequency energy concentration. Then, the speech signal to be enhanced (obtained from source enhancement) is passed through the modeled LP filter which results in speech with enhanced spectral peaks and with more energy towards higher frequency. This is evident from Figure 6, where, all the formant peaks in the enhanced spectrum are sharpened and concentration of en-

ergy towards higher frequency region increase. Here, the source enhanced speech as described in Section 3, is passed through the formant enhancement process. Figure 7 shows the wide-band spectrogram (framesize $5ms$) of (a) normal speech, (b) after source enhancement is performed, (c) after both source and formant enhancement (d) Lombard speech for the same utterance. In the dotted regions, the enhancement can be clearly observed if all the four cases are compared.

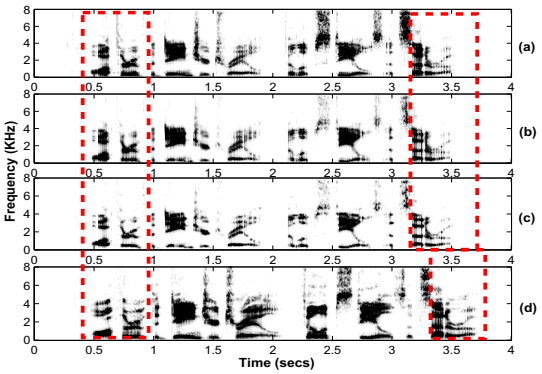

Figure 7: *Wide-band spectrogram for (a) Speech produced in quiet, (b) Strength of excitation Enhanced speech, (c) Source and formant enhanced speech (d) Lombard speech for the same utterance.*

Based on the application of TTS in practical environment, the goal of this work is to enhance synthesized speech. The synthesizer used in this case may be concatenative synthesis based on unit selection algorithm (USS) or statistical parametric speech synthesis (SPSS). As CMU_SIN database is specifically designed for application in USS based TTS in noise, in the later section of the paper, experiments are performed over synthesized speech obtained from USS based TTS using Festival framework [24]. Nevertheless, the same enhancement of strength of excitation and formant prominence is applicable to utterances synthesized using SPSS.

5. Experimental Evaluation

For evaluating the effectiveness of the proposed method, two USS based TTS systems developed using CMU_SIN database are employed. One is using speech produced in quiet (TTS_1) and the other is using Lombard speech (TTS_2). Firstly, enhancement of strength of excitation is performed as described in Section 3 over synthesized speech from TTS_1 ($ENH_1\_TTS_1$). These enhanced speech files are fed to formant enhancement process, which are termed as $ENH_2\_TTS_1$. All these four types of speech files are added with babble noise and factory noise, at different signal-to-noise ratios (SNR) (2dB, 10dB, 20dB) and the intelligibility is evaluated in terms word accuracy rate (WAR) and intelligibility based mean opinion score (MOS) over a 5 point scale, where 1 is for least intelligibility and 5 is for required intelligibility. WAR is the percentage of words which are correctly perceived by the listeners with respect to total number of words in the synthesized speech. For evaluation of MOS, the subjects are asked to decide the score based on how much attention or effort they need to pay to perceive the synthesized speech. Utterances which require less listening effort, intelligibility score will be high for those.Total 15 subjects took part in the subjective study who are research scholars having knowledge about speech intelligibility. All four types of speech files with different types and levels of noise, are coded randomly to avoid bias towards any method. Moreover, sentences used for evaluation are non-repeating. WAR and MOS obtained are shown in Table 1 for different types and levels of noise. As the database CMU_SIN is US English and the listeners are native Indian, therefore, due to mismatch in accent, maximum WAR and MOS for TTS_2 (target Lombard synthesized speech) are 75.0% and 4.2 respectively in presence of babble noise with $20dB$ SNR; accordingly, it further reduces with the decrease in SNR. Therefore, for comparison between normal synthesized speech (TTS_1) and enhanced synthesized speech $ENH_1\_TTS_1$, the gain of WAR in dB is shown in Figure 8(a). It can be observed that the gain increases with decrease in SNR and it is more useful in case of babble noise. Same can be interpreted from Figure 8(b) which depicts the gain in WAR due to the strength of excitation enhancement. A significant gain is obtained from enhancement of strength of excitation which can be observed from Figure 8(b).

Table 1: *WAR% and MOS result for babble noise and factory noise at SNR 2dB, 10dB and 20dB*

Word accuracy rate (%)					
Noise type	**Noise level (SNR)**	**TTS_1**	**$ENH_1\_TTS_1$**	**$ENH_2\_TTS_1$**	**TTS_2**
Babble noise	**2dB**	23.6	**34.8**	**38.2**	**35.7**
	10dB	55.3	58.6	61.6	63.4
	20dB	62.7	63.5	68.3	75.0
Factory noise	**2dB**	21.3	**32.0**	**38.0**	**34.9**
	10dB	43.6	52.8	59.5	61.3
	20dB	51.5	53.2	58.0	65.0
MOS for intelligibility					
Noise type	**Noise level**	**TTS_1**	**$ENH_1\_TTS_1$**	**$ENH_2\_TTS_1$**	**TTS_2**
Babble noise	**2dB**	2.2	**2.5**	**3.3**	**3.2**
	10dB	2.8	2.9	3.1	3.8
	20dB	4.0	4.0	4.1	4.2
Factory noise	**2dB**	2.3	**2.4**	**2.6**	**2.8**
	10dB	2.2	2.7	3.2	3.6
	20dB	3.2	3.3	3.5	3.9

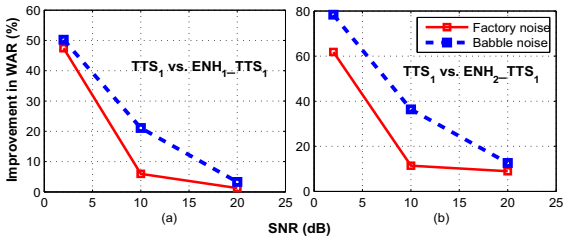

Figure 8: *Relative improvement in WAR of (a) normal speech and source enhanced speech, (b) normal speech and source and formant enhanced speech*

6. Conclusions

This work focuses on improving intelligibility of synthesized speech in presence of noise. In this regard, strength of excitation of Lombard speech and normal speech are compared and observed that, it is high in case of Lombard speech. Therefore, strength of excitation of synthesized speech is enhanced to improve intelligibility. Further, for the source enhanced speech, spectral prominence is also improved to achieve required level of intelligibility in noisy environment. Future work may focus on other aspects of the speech signal to be enhanced for robust speech synthesis.

7. Acknowledgements

This work is funded by ongoing project on the “Development of Text-to-Speech Synthesis for Assamese and Manipuri languages” funded by TDIL, DEiTy, MCIT, GOI.

8. References

[1] W. Van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[2] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *INTERSPEECH*, 2011, pp. 1837–1840.

[3] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3261–3275, 2008.

[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] Y. Lu 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.

[6] J. H. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech communication*, vol. 20, no. 1, pp. 151–173, 1996.

[7] J. H. 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 Transactions on Audio, Speech, and Language Processing,*, vol. 17, no. 2, pp. 366–378, 2009.

[8] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Lœvenbruck, "An acoustic and articulatory study of Lombard speech: Global effects on the utterance," in *INTERSPEECH*, 2006.

[9] R. J. Niederjohn and J. H. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Transactions on Acoustics, Speech and Signal Processing,*, vol. 24, no. 4, pp. 277–282, 1976.

[10] M. D. Skowronski and J. G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Communication*, vol. 48, no. 5, pp. 549–558, 2006.

[11] S. D. Yoo, J. R. Boston, A. El Jaroudi, C. C. Li, J. D. Durrant, K. Kovacyk, and S. Shaiman, "Speech signal modification to increase intelligibility in noisy environments," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1138–1149, 2007.

[12] T. C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[13] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Mel cepstral coefficient modification based on the glimpse proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise," *INTERSPEECH*, 2012.

[14] C. Valentini-Botinhao, E. Godoy, Y. Stylianou, B. Sauert, S. King, and J. Yamagishi, "Improving intelligibility in noise of HMM-generated speech via noise-dependent and-independent methods," in *ICASSP*. IEEE, 2013, pp. 7854–7858.

[15] C. Valentini-Botinhao, J. Yamagishi, S. King, and Y. Stylianou, "Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMM-based synthetic speech in noise." in *INTERSPEECH*, 2013, pp. 3567–3571.

[16] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Analysis of HMM-based Lombard speech synthesis." in *INTERSPEECH*, 2011, pp. 2781–2784.

[17] B. Langner and A. W. Black, "Creating a database of speech in noise for unit selection synthesis," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[18] G. K. Anumanchipalli, P. K. Muthukumar, U. Nallasamy, A. Parlikar, A. W. Black, and B. Langner, "Improving speech synthesis for noisy environments." in *SSW*, 2010, pp. 154–159.

[19] T. Drugman and T. Dutoit, "Glottal-based analysis of the Lombard effect." in *INTERSPEECH*, 2010, pp. 2610–2613.

[20] G. Bapineedu, B. Avinash, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of Lombard speech using excitation source information." in *INTERSPEECH*, 2009, pp. 1091–1094.

[21] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *The Journal of the Acoustical Society of America*, vol. 126, no. 4, pp. 2061–2071, 2009.

[22] K. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1602–1613, 2008.

[23] A. A. Reddy, N. Chennupati, and B. Yegnanarayana, "Syllable nuclei detection using perceptually significant features," in *INTERSPEECH*, 2013.

[24] P. Taylor, A. W. Black, and R. Caley, "The architecture of the Festival speech synthesis system," in *Proceedings of the Third ESCA Workshop in Speech Synthesis*, 1998, pp. 147–151.