Enhancing Speech Intelligibility Based on Noise Characteristics

Mayur Jagtap and Preeti Rao
Department of Electrical Engineering, Indian Institute of Technology Bombay, India
Email: {mnjagtap, prao}@ee.iitb.ac.in

***Abstract*—In degraded listening conditions, speakers are known to adapt their speech via the Lombard reflex to make it more comprehensible. This characteristic has been used in previous work to modify speech recorded in quiet before it is rendered in a noisy environment. The spectral modifications used have been found to be effective in low-pass noise such as babble noise. In this work, we investigate intelligibility enhancement of speech in completely different noise characteristics, namely aircraft noise, with its dominant high-frequency components. Natural Lombard speech elicited in aircraft noise was observed to be spectrally similar to Lombard speech in babble noise and showed no intelligibility benefit in a listening test in the presence of aircraft noise. Synthetic modifications using a data dependent optimization based on a perceptual measure are investigated to obtain intelligibility enhancement in aircraft noise.**

I. Introduction

Speech communication is an essential part of our day to day life and most of the time we don't have any control on the surrounding environment while realizing it. Humans exposed to such an environment adapt their speech characteristics to make it more comprehensible to a listener. The obvious strategy followed by humans is increasing the speech energy but along with that they adapt the speech spectral shaping as well as supra-segmental properties. This phenomenon, which is known as Lombard reflex, has been studied by many researchers over these years [1], [2], [3]. It has been observed that Lombard speech is more intelligible as compared to normal speech even when the same signal to noise energy ratio is maintained [3]. Different aspects of Lombard speech such as spectral modification, temporal or prosodic modification, energy distribution throughout the speech and phonetic modification have been analysed in past literature. Even though certain aspects such as spectral tilt have been observed by most, the robustness of such aspects for different noises is unclear.

Lu and Cooke [4] studied effect of reduced spectral tilt on the intelligibility of normal speech. They found that reducing the spectral tilt of normal speech to match that of Lombard speech significantly increases the intelligibility in babble noise. Zorila et al. [5] implemented similar method, of filtering for spectral flattening along with a dynamic range compression (SSDRC) which improved the intelligibility further. Jokinen et al. [6] combined spectral tilt flattening with formant sharpening to prove the intelligibility enhancement. But they did not comment on performance of individual methods. Godoy et al. [7] analyzed effect of expansion of vowel space on speech intelligibility and observed that vowel space expansion did not contribute to the intelligibility. All these studies analyzed Lombard speech characteristics in either babble noise or speech shaped noise and proposed the enhancements.

978-1-4799-6619-6/15/$31.00

This work studies the characteristics of Lombard speech in two distinct noise types: babble and aircraft noise. Lu and Cooke [8] did a similar study in low-pass filtered and high-pass filtered noises but they did not evaluate the subjective intelligibility of the speech in the two noises. We have also investigated a data-driven spectral enhancement technique based on optimization of a perceptual measure, known as the glimpse proportion measure, to improve intelligibility of normal speech in aircraft noise. The corpus used for above analysis was motivated by a real-world application, namely public announcements where pre-recorded audio is played out in possibly noisy environment. We use a previously designed speech database comprising railway announcements in Marathi [9].

The organization of the paper is as follows. Section II explains the database used in this study. Section III evaluates subjective intelligibility of collected data. Spectral modification technique enhancing the intelligibility of speech is proposed in Section IV. Section V discusses the results of intelligibility evaluation performed on proposed speech modification method while Section VI presents conclusion and future work.

II. Speech database

The Lombard speech was recorded in two noise types: babble and aircraft noise. Following section explains their characteristics.

A. Characteristics of noises used for data-collection

The noises chosen for the study are babble noise and aircraft noise taken from NOISEX-92 database [10]. Babble noise has most of the energy concentrated in lower part of frequency spectrum (below 1 kHz) where as aircraft noise has significant energy in the higher formant regions (2 - 3 kHz) as well as in the lower frequency regions. Figure 1 shows the power spectral density of aircraft noise and babble noise.

B. Corpus collection

The corpus used in this study was recorded from 4 native Marathi speakers. Each speaker recorded 3 sets of 10 sentences in different noise conditions viz. no noise, aircraft noise and babble noise. The procedure followed for corpus collection was same as that of [9]. The equivalent English translation of the template used for data collection is *"Train number **A B C D** down, **NAME** Express, will halt at platform number **Y** instead*

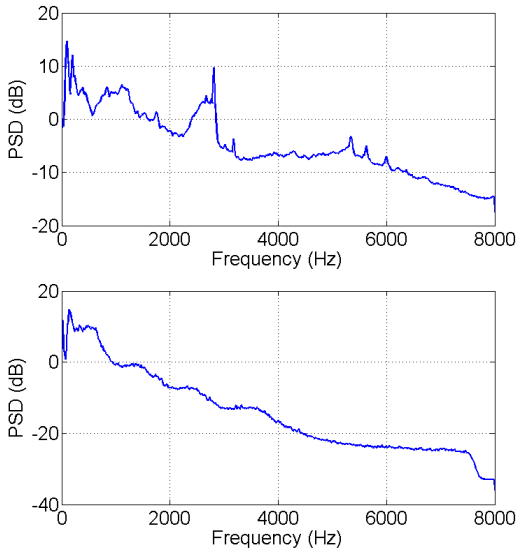

Fig. 1. Power spectral densities (a)aircraft noise (top) (b)babble noise (bottom), Estimated using periodogram averaging for frame size of 64 msec (1024 sample for 16 kHz sampling rate) without overlapping

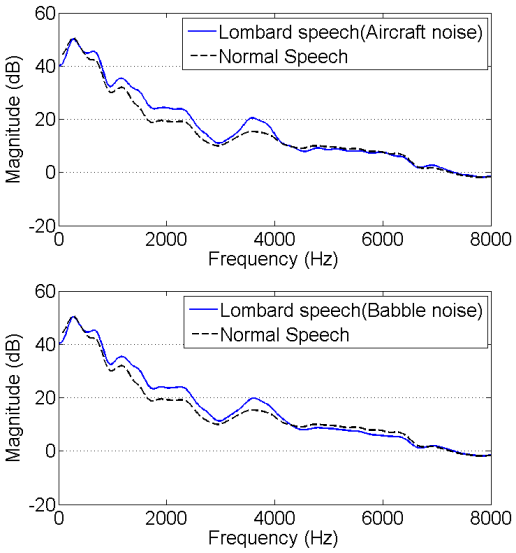

Fig. 2. Long term average spectra of Lombard speech of speaker 4 in aircraft noise (top) and babble noise (bottom) compared with normal speech

TABLE I. LIST OF WORDS USED FOR DATABASE COLLECTION

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)

*of platform number **X**"*. As depicted in Table I train number A, B, C, D were digits from 0 to 9, platform numbers X and Y were numbers from 10 to 19 and 5 different express train names were used. A set of 10 sentences was designed using above template. The sentences were formed in such a way that each digit from 0 to 9 occurred once in each of four possible positions. Same strategy was followed for platform numbers. Each train name from the list of five was repeated once.

While recording, babble noise was played through SONY MDR-XD200 headphones to the human speaker. The speaker was asked to read 10 sentences in format specified above while listening to the noise. No sound measuring equipment was available for the test. Thus for a relative calibration of the sound level two sets of 10 sentences were recorded in two levels of noise 30 dB apart. Lower level of noise was barely perceptible and thus the recordings done in this noise level were assumed to be normal speech. The recording was done using SHURE SM58 microphone at 16 kHz. The same set-up was maintained throughout all the recordings. Same procedure was followed for recording Lombard speech in aircraft noise.

III. SPECTRAL ANALYSIS AND INTELLIGIBILITY EVALUATION OF RECORDED DATA

Spectral analysis of data was done by computing the long-term average spectra of each speech type produced by each speaker. Godoy et al. [7] have used the same method for analyzing energy distribution of the speech in the frequency domain.

A. *Spectral analysis*

To calculate the average spectrum of the speech first the speech was normalized to have unit RMS energy. Then cepstrally smoothened speech spectra were computed at 10 msec frame intervals using 20 msec Hamming window and the first 40 cepstral coefficients. Finally frame-wise spectra were added together and then divided by the number of frames to compute the average spectrum of the speech. This was done for each of the speakers for all three categories of speech viz. normal speech, Lombard speech in aircraft noise and Lombard speech in babble noise.

Figure 2 shows long term average spectra computed for Speaker 4. It can be observed that average spectra of Lombard speech in both noises were comparable. Speakers followed same strategy of boosting speech energy in 500-4500 Hz region in both noises. This observation was in agreement with [8], [7].

B. *Intelligibility evaluation through subjective listening test*

Subjective intelligibility of recorded data was evaluated through listening tests. Two listening tests, one each for aircraft and babble noise were performed by the listeners. Each listener listened to the sentences formulated from the recorded data by randomizing the train number, name and platform number. Words uttered at different positions in a sentence will have different prosodies. To account for this, while randomizing, the positions of train and platform numbers were maintained fixed to the specific slot in which it was recorded. Listening test consisted a total of 20 sentences, 10 Lombard and 10 normal. In each of the sentences the listener was supposed to identify four digits of train number, train name and the two platform numbers. This setup was same as the one used in [9].

TABLE II. RESULTS OF LISTENING TESTS (% WER)

	babble noise			aircraft noise		
	SNR (dB)	Normal speech	Lombard speech	SNR (dB)	Normal speech	Lombard speech
Train no.	-3 -6	21 56	5 20	-8 -11	19.7 48.3	20.5 37.5
Platform no.	-3 -6	14 25	5 12.5	-8 -11	29.5 40.6	29 37
Train name	-3 -6	0 4	4 0	-8 -11	4 4	0 8

To compare the intelligibility gains in the two noises, it is necessary that the word error rate of normal speech in both noises is comparable too. Therefore we used different SNRs for the two noises. Table II summarizes the result of listening test. The listening test was performed by 5 listeners, all speaking Marathi as their mother tongue. From the results it is clearly observed that WER reduces for Lombard speech in babble noise. This was expected. Similar results have been observed in previous literature [2], [1]. But for aircraft noise, except for train numbers accuracy at -11 dB SNR, the WER of Lombard speech was comparable to that of normal speech. This was an interesting observation which was not reported in previous literature.

C. Intelligibility evaluation through objective measure

For evaluating the intelligibility of recorded data we have used glimpse proportion measure (GPM) as an objective intelligibility measure. GPM is a perceptual measure introduced by Cooke in 2006 [11]. This measure predicts intelligibility by comparing spectrotemporal excitation pattern (STEP or spectrogram) of speech and noise. The concept behind GPM is that speech perception is associated with the area of the speech spectrum least affected by noise.

Glimpses are the spectrotemporal regions where speech energy dominates noise atleast by 3 dB. GPM counts the number of speech glimpses that occur in the mixture of speech and noise. More the number of glimpses more is the GPM and so is the predicted intelligibility. Following mathematical equation expresses the GPM.

$$GPM = \frac{100}{TF} \sum_{t=1}^{T} \sum_{f=1}^{F} \mathcal{L}(S_{t,f} - (N_{t,f} + \alpha)) \qquad (1)$$

Here $S_{t,f}$ and $N_{t,f}$ are STEP or spectrogram of clean speech and noise respectively. α is the threshold above which the glimpses are counted (in this case 3 dB). T and F are number of time and frequency channels in the speech. The operator $\mathcal{L}(.)$ counts the number of non-zero values. Thus glimpse proportion score represents the percentage area of the spectrogram where local SNR is above threshold α.

Table III shows the glimpse proportion measure results compared with subjective listening tests results. The spectrogram used while computing GPM score was computed using 50 msec window, 10 msec hop size and 1024 point FFT. GPM score was calculated from all the utterances of speaker 4 in the respective category (i.e. normal speech, Lombard speech in babble noise and Lombard speech in aircraft noise). This is compared with % word error rate of only train numbers in subjective listening tests. The platform numbers are two syllabic and few of them have one common syllable creating confusion. The train name accuracy is high since all the train names are 3 syllabic and the listener has to choose from only 5 train names. Thus the train number WER is the most reliable representative of intelligibility.

TABLE III. COMPARISON OF GPM SCORE WITH RESULTS OF LISTENING TESTS % WER FOR TRAIN NUMBERS ONLY

	babble noise		
	SNR(dB)	GPM Score	% WER
Normal speech	-3 -6	6.17 4.22	21 56
Lombard speech (babble noise)	-3 -6	7.25 5	5 20
	aircraft noise		
	SNR(dB)	GPM Score	% WER
Normal speech	-8 -11	1.11 0.71	19.75 48.3
Lombard speech (aircraft noise)	-8 -11	1.34 0.85	20.5 37.5

For babble noise difference between GPM score of normal speech and Lombard speech is evident and the same trend is reflected in subjective test results (% WER). For aircraft noise GPM score is very small. The correlation of GPM score and WER for aircraft noise is evident from the values corresponding to -8 dB and -11 dB SNR. Higher the GPM score lower is the WER. The difference between the GP scores of normal and Lombard speech is very small, around 0.1. The trend is maintained for WER but it is not as evident as for babble noise. Thus GPM score captured relative trend in intelligibility.

IV. SPECTRAL MODIFICATION: MAXIMIZING GLIMPSE PROPORTION MEASURE (GPM)

Tang and Cooke (2012) [12] have proposed a noise-dependent intelligibility enhancement method. We have used same method for speech enhancement in aircraft noise. The method obtains a spectral filter which will maximize the GPM score of the speech in given noise. Figure 3 shows the block diagram of the implementation. Our aim is to find a spectral filter (W = $\{w_0, w_1, ..., w_N\}$) which maximizes GPM score of normal speech (s(t)) in noise (n(t)) under the constraint that input and output energies are constant. This optimization problem can be expressed as follows.

$$\begin{aligned} \underset{W}{\text{maximize}} \quad & \text{GPM}\{s(t) \star w(t)\} \\ \text{subject to} \quad & \sum_{t=1}^{T} \{s(t) \star w(t)\}^2 = \sum_{t=1}^{T} \{s(t)\}^2 \end{aligned} \qquad (2)$$

where, function GPM(s(t)) computes GPM score of s(t) in noise n(t) using Equation 1. w(t) is time domain representation of spectral filter W.

To solve above optimization problem a genetic algorithm [13] was used. Genetic algorithm works efficiently in higher dimensional optimization problems. The terminologies of genetic algorithm in context of our optimization problem are explained in following section.

A. Genetic algorithm

Genetic algorithms (GA) is large class of optimization algorithms which is inspired from the concepts in evolution

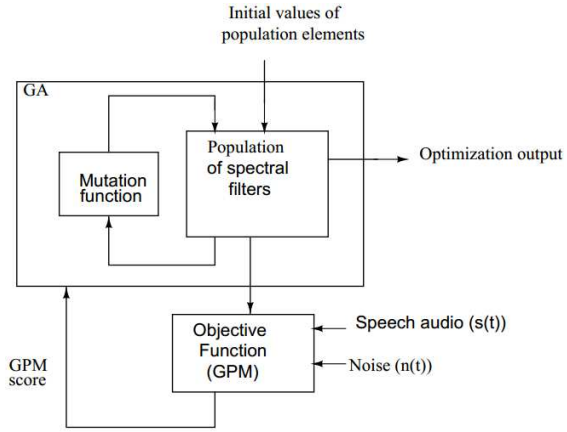

Fig. 3. Block diagram of genetic algorithm implementation

TABLE IV. GENETIC ALGORITHM PARAMETERS

constraint	boosting bounds	[-50 50] dB
population	size	300
	initial variable value	0
	elite count	30
stopping criteria	generations	500
	stall generations	5

process of animals such as mutation, inheritance, selection and crossover. The details of genetic algorithm can be found in [13]. This section tries to motivate the optimization procedure followed in genetic algorithm. Genetic algorithm starts with a set of candidate solutions to the optimization problem. The set is called *population*. In our problem the population consists of various spectral weights. The spectral weights from the population are modified in each iteration. This process is called *evolution* and population in each iteration is called *generation*.

In the beginning the population is either initialized with a fixed valued set or chosen randomly. Then iteratively the population is updated from one generation to next generation. The objective function in the optimization problem is called as *fitness function*. In one iteration each element in the population set evaluated using the fitness function i.e. in our case GPM score is computed using each spectral weight as specified in Equation 2. The elements with highest value of objective function in population set are identified and are mutated to form next generation. Here modification of spectral weights is called as *mutation*. In this way the process continues. There are different categories of genetic algorithms depending on mutation function used [14].

B. Optimization of GPM using genetic algorithm

Figure 3 shows the block diagram of implementation of genetic algorithm. In the optimization process, all the normal speech data recorded from speaker 4 (10 sentences) and aircraft noise section of 3 min 55 sec is given as input to objective function block. The GA block represents genetic algorithm command 'ga' from Global Optimization Toolbox in MATLAB. Thus GA block searches for spectral weights which maximizes output of objective function.

The population specified in GA block consists of possible spectral filters which are specified in dB scale. The population consisted of such 300 instances of spectral filters W_is each with 512 dimensions. We first initialized the spectral filters with 0 and then started the optimization. In each iteration all the members of the population were evaluated using objective function (GPM). We wanted to evaluate the GPM on normal speech in aircraft noise after applying the spectral filter. Thus normal speech of speaker 4 and aircraft noise were given as fixed input to GPM block. GPM score was evaluated using the inputs after the application of spectral filter on the speech. This score was returned to the genetic algorithm. After evaluating all the members of population, the members with highest fitness (i.e. highest value of objective function) were identified as most likely solutions to the optimization problem. Next population was formed using these members and their modified versions and next iteration was started. In this way the genetic algorithm progresses towards the optimum solution.

The optimization stops when either objective function value of the best individual of the population remains constant for specified number of stall generations or the number of generation reaches the maximum limit. Here stall generations were set to 5 and maximum generation to 500. Table IV summarizes the parameters used in genetic algorithm. The population size was chosen to be 300. Population size should not be too low. This reduces the probability of finding optimum solution. But we can't keep it too high due to memory constraint. Also it takes more time to evaluate larger population. Elite count was 30 i.e. 10% of total population. This specifies number of candidates selected to form next generation after mutation.

To calculate GPM score first the speech and noise files are loaded. Then spectrogram of both is calculated using using 50 msec window and 10 msec hop-size. Then spectral filter is added to the speech spectrogram in dB scale. Finally using Equation 1 GPM score is evaluated.

V. RESULTS AND DISCUSSION

The population used in this method was initialized with 0 values. Each element of the population was an array with 512 values of spectral filter. Glimpse proportion measure was the objective function. The maximum generation stopping criteria was 500, thus after 500 generations the solution of the optimization was obtained. Figures 4 shows the frequency domain filters obtained after optimization in aircraft noise and babble noise. Note that spectral filter in aircraft noise enhances spectral energy between 1-5 kHz and attenuates the frequencies below 1 kHz. This constant filter was used for enhancing the intelligibility in aircraft noise. For comparison we also obtained similar filter for babble noise. It shows a similar characteristics, amplifying all the frequencies above 2 kHz. The difference between the two is observed above 5 kHz, where gain of filter for aircraft noise decreases and gain of filter for babble noise increases. Figure 5 shows the long term spectra of normal speech before and after filtering using the filter designed for aircraft noise.

Table V shows the results of listening test performed by 5 listeners after applying proposed spectral modification. The proposed method did not show any significant improvements in the intelligibility. Spectral modification method worsened the WER.

Table VI shows the comparison of train number WER and GPM score. The spectral modification increased the GPM

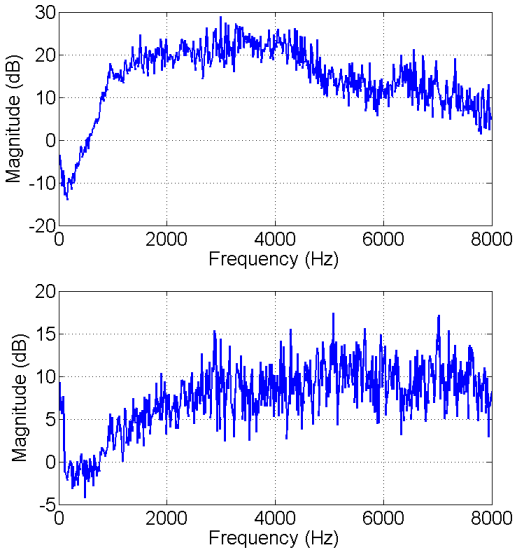

Fig. 4. Filter frequency response obtained after optimization of GPM in aircraft noise (top) and babble noise (bottom)

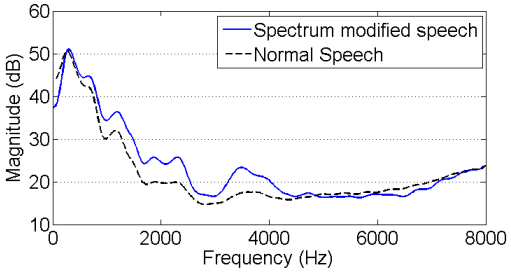

Fig. 5. Comparison of normal speech filtered using obtained filters for aircraft

TABLE V. RESULTS OF SUBJECTIVE LISTENING TESTS PERFORMED ON NORMAL SPEECH, LOMBARD SPEECH AND SPEECH MODIFIED WITH PROPOSED METHOD IN AIRCRAFT NOISE(% WER)

	SNR(dB)	Normal speech	Lombard	Spec modified
Train no.	-8	19.7	20.5	38
	-11	48.3	37.5	43
Platform no.	-8	29.5	29	30
	-11	40.6	37	56
Train name	-8	4	0	4
	-11	4	8	28

TABLE VI. COMPARISON OF RESULTS OF SPEECH MODIFICATION METHOD FOR AIRCRAFT NOISE FOR ONLY TRAIN NUMBERS WITH GPM SCORE

	SNR(dB)	GPM score	% WER
Normal speech	-8	1.1	19.7
	-11	0.71	48.3
Natural Lombard speech	-8	1.34	20.7
	-11	0.85	37.5
Spectral filtering	-8	1.39	38
	-11	0.86	43

score of the speech as expected, but the WER of spectral modification was worse than normal speech in -8 dB SNR whereas there was a slight improvement in WER for -11 dB SNR.

VI. CONCLUSION AND FUTURE WORK

This work investigated spectral characteristics and intelligibility of Lombard speech in babble and aircraft noise which have completely different spectral properties. Subjective listening test results showed that intelligibility of Lombard speech is only comparable to (not better than) normal speech in aircraft noise. It was observed that speakers follow an articulation strategy of reducing spectral tilt irrespective of noise characteristics. This is in agreement with the observations made by [8]. This strategy was effective in babble noise since it increases the local spectral SNR where it is most effective. But in aircraft noise, which has distributed spectral energy, this strategy of reducing spectral tilt fails to provide any intelligibility gain.

A previously proposed data-dependent spectral modification technique was investigated to maximize the Glimpse Proportion Measure. While the GPM did indeed improve, our subjective listening tests did not evidence any improvement in intelligibility. After the modification, the speech sounded synthetic which was also observed by Tang and Cooke [12]. It is possible that the spectral shaping was influenced mostly by vowel characteristics due their temporal dominance in the speech signal. The consonants which matter most to speech intelligibility may not have benefited from the overall increase in GPM. Future work will focus on modifying the spectral optimization to achieve improved intelligibility in aircraft-type noise and also implement speech dependent enhancement filter. For further experimentation we may also try and simulate real life environment by playing audio through loud-speakers while collecting data.

ACKNOWLEDGMENT

The authors would like to thank all the participants of audio recordings and listening tests. We would also like to thank Hitesh Tulsiani and V. Shrikant for valuable inputs.

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] J. C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[3] 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.

[4] 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.

[5] T. C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Interspeech*, USA, 2012.

[6] 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.

[7] 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.

[8] Y. Lu and M. Cooke, "Speech production modifications produced in the presence of low-pass and high-pass filtered noise," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1495–1499, 2009.

[9] N. Jathar and P. Rao, "Acoustic characteristics of critical message utterances in noise applied to speech intelligibility enhancement," in *Interspeech*, Singapore, 2014.

[10] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[11] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[12] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement." in *Interspeech*, USA, 2012.

[13] J. H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975.

[14] M. Mitchell, *An Introduction to Genetic Algorithms*. MIT press, 1998.