INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Transfer Learning for Speech Intelligibility Improvement in Noisy Environments

Ritujoy Biswas[1], *Karan Nathwani*[1], *Vinayak Abrol*[2]

[1] Indian Institute of Technology Jammu, India
[2] Indraprastha Institute of Information Technology Delhi, India

ritujoy.biswas@iitjammu.ac.in, karan.nathwani@iitjammu.ac.in, abrol@iiitd.ac.in

Abstract

In a recent work [1], a novel Delta Function-based Formant Shifting approach was proposed for speech intelligibility improvement. The underlying principle is to dynamically relocate the formants based on their occurrence in the spectrum away from the region of noise. The manner in which the formants are shifted is decided by the parameters of the Delta Function, the optimal values of which are evaluated using Comprehensive Learning Particle Swarm Optimization (CLPSO). Although effective, CLPSO is computationally expensive to the extent that it overshadows its merits in intelligibility improvement. As a solution to this, the current work aims to improve the Short-Time Objective Intelligibility (STOI) of (target) speech using a Delta Function that has been generated using a different (source) language. This transfer learning is based upon the relative positioning of the formant frequencies and pitch values of the source & target language datasets. The proposed approach is demonstrated and validated by subjecting it to experimentation with three different languages under variable noisy conditions.

Index Terms: Speech Intelligibility, Transfer Learning, Formant Ratio, Formant Shifting, Pitch Ratio

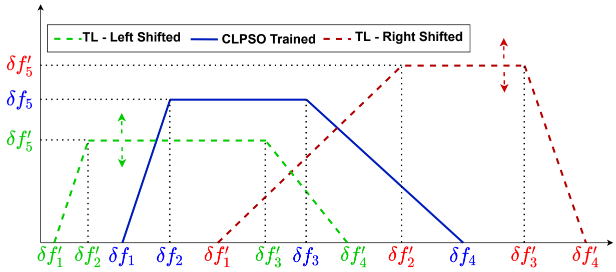

Figure 1: *Delta Function obtained through CLPSO and Transfer Learning (TL) via left (L) and right (R) shifting.*

1. Introduction

The prime target of most, if not all, speech processing efforts made in the literature, has been the improvement in the quality of speech. The intent, generally was to mitigate the effect of background noise. The obvious idea behind such approaches was the improvement in SNR that ultimately led to an improvement in quality [2–11]. However, often the performance of existing enhancement methods does not target the improvement of intelligibility while addressing the quality. This may be attributed to the fact that most speech enhancement algorithms use a cost function that does not correlate well with intelligibility [12, 13]. Further, intelligibility is guided by articulation of the uttered speech and, thereby, depends upon the various parts of speech such as vowels, consonants, plosives, etc. It is therefore, plausible to say that variations in these parts of speech lead to variation in manner of articulation and place of articulation, and are not well correlated with quality improvement.

Over the past decade, various studies have proposed approaches to directly address the problem of improving speech intelligibility. Some of the earliest methods involved boosting of higher frequencies while dealing with low pass noise [14] and dynamic range compression [15]. Subsequent studies include developing an Optimum Linear time-invariant Filter (OLF) which maximizes the Speech Intelligibility Index (SII) [16, 17], or by optimally Redistributing Energy (RE) over time and frequency [18, 19]. However, despite the efficacy, their performance degrades at very low SNRs. To address these issues, the authors in [13, 20, 21] proposed an approach to improve the intelligibility by shifting the formants away from the region of noise using a novel delta function (see Fig. 1). Originally, the parameters of this delta function were chosen empirically which resulted in a sub-optimal performance due to various auditory artifacts [20]. To rectify such problems, a combined modification was proposed in [13]. To further improve the results that were initially obtained through trial and error, a meta-heuristic search algorithm was implemented namely Comprehensive Learning Particle Swarm Optimization (CLPSO) [22–24], in [1, 25–27]. This algorithm maximizes the Short Time Objective Intelligibility (STOI) and produces the optimum shaping parameters for the delta function. The added benefit of this approach was that there was no requirement of additional smoothing after formant shifting and *apriori* noise statistics as required in [10, 13, 20].

Although the CLPSO based formant shifting approach works well in practice, it is computationally and temporally very expensive pertaining to the long time taken by the genetic algorithm to converge to the optimal parameters. Furthermore, since the parameters were optimized for a single language only, the model cannot be expected to be equally robust for other languages. Addressing the aforementioned issue, the proposed approach in this paper attempts to shift the formants in real time away from regions where the intelligibility is affected by noise, by modifying an already generated Delta Function (see Figure 1). This is coupled with a real-time modification in scale of delta function using pitch. The direction of these shifts will depend upon the comparative examination of the formant frequencies and pitch values of the source and target languages.

The rest of the paper is organized as follows: Section 2 briefs the CLPSO based formant shifting followed by the proposed transfer learning mechanism in section 3. Section 4 evaluates and validates the results. Section 5 concludes this paper.

http://dx.doi.org/10.21437/Interspeech.2021-150

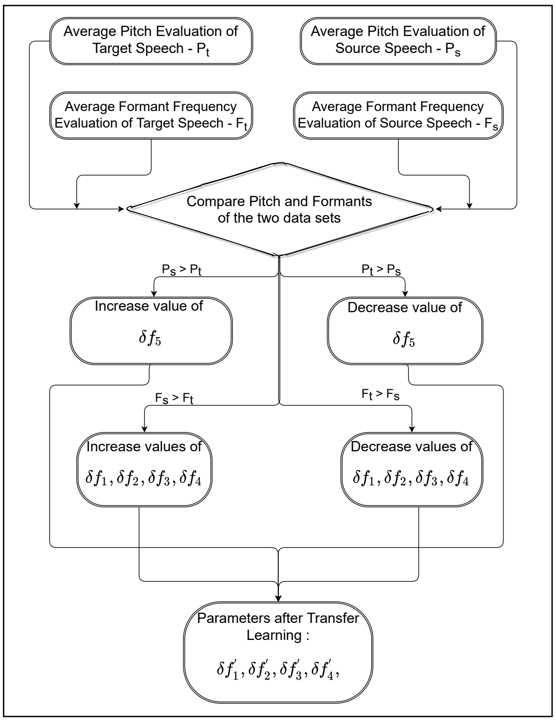
Average Pitch Evaluation of Target Speech - P_t
Average Pitch Evaluation of Source Speech - P_s
Average Formant Frequency Evaluation of Target Speech - F_t
Average Formant Frequency Evaluation of Source Speech - F_s
Compare Pitch and Formants of the two data sets
$P_s > P_t$
$P_t > P_s$
Increase value of δf_5
Decrease value of δf_5
$F_s > F_t$
$F_t > F_s$
Increase values of $\delta f_1, \delta f_2, \delta f_3, \delta f_4$
Decrease values of $\delta f_1, \delta f_2, \delta f_3, \delta f_4$
Parameters after Transfer Learning : $\delta f_1', \delta f_2', \delta f_3', \delta f_4',$

Figure 2: *Block Diagram for the Transfer Learning*

2. Review : CLPSO based Formant Shifting

To begin with, the Delta Function parameters (as shown in Figure 1) are evaluated using a meta heuristic search approach- CLPSO. This training is done on the data sets of three languages - English (CHAINS [28]), French (HINT [29]) and German (VOXFORGE [1]). In essence, CLPSO is a variant of particle swarm optimization which operates on a *swarm* of search agents (*particles*). There are four attributes to each particle : *position* ($\mathcal{X}$), *velocity* ($\mathcal{V}$), *personal best position* ($\hat{\mathcal{X}}$) and *swarm best position* ($\mathcal{X}^*$). The ($\mathcal{X}$) represents a vector of candidate solution, *i.e.*, $\mathcal{X} = \{x_1,\ x_2,\ \ldots,\ x_D\}$, where '$D$' represents the number of search variables/dimensions. For example, in this work, the i^{th} particle explains the candidate set of shaping parameters given by $\mathcal{X}_i$ with $D = 5$. The first four shaping parameters $\delta f_1, \delta f_2, \delta f_3, \delta f_4$ represent the spectral locations of the edges, while δf_5 gives the height of the delta function (the maximum value of shift in frequency).

During the CLPSO [23, 27] search progress, each particle stores the best position found by itself ($\hat{\mathcal{X}}$) and the best position found by the entire swarm ($\mathcal{X}^*$). These positions are updated with each iteration. The particles co-operate with each other throughout the search process by sharing the information about *personal best*. In each iteration, the particles move on the search landscape till the optimum swarm best position is obtained. Following the guidelines in [22], the search parameters of CLPSO are chosen. The resulting shaping parameters obtained under different noise conditions are given in Table 1. Following the optimization, the shift in each formant is controlled by these five *'shaping'* parameters as:

Table 1: *Identified Shaping Parameters for Babble, Car (50 and 130 mph) and Train (50mph) at multiple languages using CLPSO*

Language	Noise	SNR	Parameters				
			δf_1	δf_2	δf_3	δf_4	δf_5
English	Babble	-26	29.69	245.37	2368.44	3709.58	498.34
		-14	0.83	151.04	1678.33	2545.80	359.45
		-8	2.80	157.51	1713.87	2598.37	317.94
	Car$_{50}$	-26	33.67	244.75	1681.41	3529.69	473.70
		-14	10.62	169.65	1562.52	2417.89	302.78
		-8	16.98	168.28	1621.15	2460.97	273.93
	Car$_{130}$	-26	0.61	168.10	1549.85	2822.38	339.78
		-14	0.21	653.60	1741.20	3888.31	491.58
		-8	31.17	946.16	1566.19	3407.95	498.49
	Train$_{50}$	-26	42.99	316.55	2332.73	3971.02	499.83
		-14	41.45	249.15	1543.05	2962.60	436.11
		-8	33.71	192.70	1740.59	2495.88	305.17
French	Babble	-26	42.69	151.78	2336.10	3713.96	466.68
		-14	1.20	150.66	1960.52	3269.08	400.83
		-8	1.01	150.43	2236.74	3998.04	393.18
	Car$_{50}$	-26	12.91	154.80	2155.89	3895.12	451.01
		-14	9.36	151.89	2117.58	3760.25	345.62
		-8	11.90	153.41	2269.78	3691.70	318.82
	Car$_{130}$	-26	1.96	153.72	2375.94	3687.42	415.64
		-14	5.78	367.45	2226.83	3882.50	476.60
		-8	5.30	886.27	2344.05	3869.75	438.21
	Train$_{50}$	-26	38.07	151.91	2130.43	3659.52	450.82
		-14	19.17	152.61	2084.00	3634.54	422.42
		-8	8.62	151.44	2020.82	3570.62	364.55
German	Babble	-26	63.86	210.18	1718.89	3121.17	499.20
		-14	47.38	252.33	2001.32	2400.29	443.79
		-8	50.26	255.90	1876.57	2405.80	422.81
	Car$_{50}$	-26	59.71	225.65	1848.17	2427.03	487.62
		-14	53.81	256.10	1961.15	2412.74	396.50
		-8	60.00	259.08	2022.84	2419.76	379.23
	Car$_{130}$	-26	39.51	269.11	1895.83	2445.06	442.74
		-14	117.43	282.88	1883.99	2472.84	457.69
		-8	144.26	283.96	1974.94	2446.33	425.56
	Train$_{50}$	-26	58.90	213.39	1579.52	3028.87	488.17
		-14	72.40	220.32	1969.56	2456.05	478.44
		-8	54.05	235.03	2006.41	2406.20	406.16

$$\hat{F} = \begin{cases} \frac{\delta f_5}{(\delta f_2 - \delta f_1)} + F, & if \quad \delta f_1 \leq F < \delta f_2 \\ \delta f_5 + F, & if \quad \delta f_2 \leq F < \delta f_3 \\ \frac{-\delta f_5}{(\delta f_4 - \delta f_3)} + F, & if \quad \delta f_3 < F \leq \delta f_4 \\ F, & otherwise \end{cases} \tag{1}$$

where, 'F' and '$\hat{F}$' respectively denote the original and shifted formants. Once the shifted formants ($\hat{F}$) have been obtained, the modified signal needs to be synthesized [13, 20]. To achieve this, the new poles $\hat{P}(f, m)$ are first obtained from $\hat{F}$ for each frame (m) and each formant frequencies (f). Thereafter, the modified Linear Predictive (LP) spectrum $\hat{A}(f, m)$ is computed. The modified LP spectrum, together with the spectrum of the original signal $A(f, m)$ is used to perform spectral masking to obtain $\hat{S}(f, m)$. The modified signal, $\hat{s}(n, m)$ is then obtained through inverse STFT on $\hat{S}(f, m)$. This is performed for all frames and the signal $\hat{s}[n]$ is constructed through Overlap and Add (OLA) synthesis. The energy is normalized to that of the original signal and subsequently, the STOI values are evaluated for clean (modified) and clean (modified) added with noise.

3. Transfer Learning for Speech Intelligibility Improvement

The primary focus of this work shall be to reduce the computational expense and complexity that results due to training and re-training the model to suit the needs of dataset of differing languages. Currently, several hours are required to optimize the Delta Function parameters at a certain SNR on a system hav-

[1] https://tinyurl.com/ydn7wt5c

Table 2: *STOI checked on French (FRE) (Self, Post Transfer Learning from English (ENG) & from German (GER))*

Noise	SNR	Self (CLPSO on FRE)			Transfer Learning (CLPSO on ENG & GER)			
		$\text{STOI}_{\text{O:O+N}}^{\text{NM }(FRE)}$	$\text{STOI}_{\text{M:M+N}}^{FRE \rightarrow FRE}$	Imp(%)	$\text{STOI}_{\text{M:M+N}}^{ENG \rightarrow FRE}$	Imp(%)	$\text{STOI}_{\text{M:M+N}}^{GER \rightarrow FRE}$	Imp(%)
Babble	-8	0.4382	0.5667	29.3253%	0.4675	**6.6857%**	0.5030	**14.7966%**
	-14	0.3149	0.4493	42.6681%	0.3446	**9.4203%**	0.3583	**13.7790%**
	-26	0.2465	0.3440	39.5703%	0.2662	**7.9988%**	0.2440	-0.9828%
Car_{50}	-8	0.7112	0.7856	10.4512%	0.7222	**1.5419%**	0.7649	**7.5497%**
	-14	0.5374	0.6590	22.6395%	0.5612	**4.4420%**	0.6146	**14.3613%**
	-26	0.3060	0.4347	42.0486%	0.3246	**6.0920%**	0.3424	**12.0399%**
Car_{130}	-8	0.7551	0.7914	4.7959%	0.7566	**0.1873%**	0.7882	**4.3799%**
	-14	0.6203	0.7009	12.9915%	0.6330	**2.0474%**	0.6778	**9.2638%**
	-26	0.3902	0.5335	36.7422%	0.4208	**7.8494%**	0.4527	**16.0278%**
Train_{50}	-8	0.4712	0.5834	23.8182%	0.4810	**2.0825%**	0.5172	**9.7587%**
	-14	0.3416	0.4574	33.9116%	0.3546	**3.7997%**	0.3699	**8.2941%**
	-26	0.2495	0.3537	41.7526%	0.2644	**5.9628%**	0.2492	-0.1159%

Table 3: *STOI checked on German (GER) (Self, Post Transfer Learning from English (ENG) & from French (FRE))*

Noise	SNR	Self (CLPSO on GER)			Transfer Learning (CLPSO on ENG & FRE)			
		$\text{STOI}_{\text{O:O+N}}^{\text{NM }(GER)}$	$\text{STOI}_{\text{M:M+N}}^{GER \rightarrow GER}$	Imp(%)	$\text{STOI}_{\text{M:M+N}}^{ENG \rightarrow GER}$	Imp(%)	$\text{STOI}_{\text{M:M+N}}^{FRE \rightarrow GER}$	Imp(%)
Babble	-8	0.4469	0.5820	30.2275%	0.4690	**4.9416%**	0.3994	-10.6446%
	-14	0.3540	0.5006	41.4246%	0.3783	**6.8627%**	0.3700	**4.5248%**
	-26	0.2776	0.4144	49.2636%	0.2956	**6.4684%**	0.3568	**28.5260%**
Car_{50}	-8	0.6402	0.7462	16.5530%	0.6555	**2.3941%**	0.5847	-8.6655%
	-14	0.5230	0.6508	24.4395%	0.5434	**3.9028%**	0.4766	-8.8701%
	-26	0.3461	0.4830	39.5760%	0.3637	**5.1042%**	0.3657	**5.6715%**
Car_{130}	-8	0.7432	0.7855	5.6842%	0.7450	**0.2312%**	0.7326	-1.4271%
	-14	0.6359	0.7061	11.0376%	0.6441	**1.2623%**	0.5758	-9.4512%
	-26	0.4133	0.5731	38.6768%	0.4411	**6.7376%**	0.4318	**4.4745%**
Train_{50}	-8	0.5067	0.6290	24.1339%	0.5189	**2.4126%**	0.4303	-15.0879%
	-14	0.4064	0.5296	30.3199%	0.4193	**3.1909%**	0.3750	-7.7113%
	-26	0.3068	0.4307	40.3929%	0.3191	**4.0147%**	0.3562	**16.1040%**

Table 4: *STOI checked on English (ENG) (Self, Post Transfer Learning from French (FRE) & from German (GER))*

Noise	SNR	Self (CLPSO on ENG)			Transfer Learning (CLPSO on FRE and GER)			
		$\text{STOI}_{\text{O:O+N}}^{\text{NM }(ENG)}$	$\text{STOI}_{\text{M:M+N}}^{ENG \rightarrow ENG}$	Imp(%)	$\text{STOI}_{\text{M:M+N}}^{FRE \rightarrow ENG}$	Imp(%)	$\text{STOI}_{\text{M:M+N}}^{GER \rightarrow ENG}$	Imp(%)
Babble	-8	0.4141	0.5479	32.2961%	0.4266	**3.0156%**	0.3923	-5.2729%
	-14	0.3137	0.4477	42.7275%	0.3618	**15.3478%**	0.3266	**4.1253%**
	-26	0.2541	0.3616	42.3129%	0.2639	**3.8610%**	0.2541	**5.7698%**
Car_{50}	-8	0.6424	0.7359	14.5523%	0.6192	-3.6151%	0.5603	-12.7759%
	-14	0.4928	0.6201	25.8352%	0.5103	**4.6511%**	0.4522	-8.2410%
	-26	0.3119	0.4245	36.1013%	0.3264	**4.6511%**	0.3009	-3.5173%
Car_{130}	-8	0.7304	0.7655	4.7944%	0.6252	-14.4067%	0.5930	-18.8210%
	-14	0.6066	0.6797	12.0513%	0.5435	-10.4048%	0.5097	-15.9698%
	-26	0.3927	0.5226	33.0764%	0.4344	10.6286%	0.4106	**4.5703%**
Train_{50}	-8	0.5048	0.6047	19.7821%	0.5068	**0.3871%**	0.4548	-9.9056%
	-14	0.4077	0.5018	23.0873%	0.4016	-1.4845%	0.3658	-10.2587%
	-26	0.3060	0.4046	32.2053%	0.2959	-3.3032%	0.2883	-5.7981%

ing Intel Xeon Processor. Further, this is specifically when 25 independent runs of CLPSO on 16 parallel cores are being carried out for a single SNR with each run is terminated after 3000 function evaluations. To address this, the proposed approach transfers the knowledge from an optimally generated delta function on one language by modifying the shaping parameters to suit the requirements of other languages with reasonable tradeoff in performance.

To this aim for a target dataset, the average pitch has been evaluated over all frames of a speech signal in addition to the average pitch value on the source dataset. Similarly, the average formant frequency has been evaluated on source and target language. This is done for each of the three datasets under different noisy environments. The underlying principal is to compare the average pitch and formant frequencies of the target speech utterance from (*Source Set* to *Target Set*). The steps for this transfer approach (outlined in Figure 2), which result in modified delta function in Figure 1, are as follows :

1. If the ratio of *Target Set* to *Source Set* average pitch (P_{av}) is greater than 1, the *height* (δf_5) of the delta function is raised for a more aggressive shift in formants. Similarly, if the value is less than 1, the *height* is reduced. This modification is done as : $\delta f_5^{'} = \delta f_5 \pm (P_{av} \times \delta f_5)$
2. If the ratio of *Target Set* to *Source Set* average formant frequency (F_{av}) is greater than 1, the entire delta function (i.e., $\delta f_1, \delta f_2, \delta f_3, \delta f_4$) is shifted to the right. Similarly, if the ratio is less than 1, the delta function is shifted to the left. The motive is to modify the parameters in a way that the delta function shifts the formants which account for intelligibility of speech, away from noisy regions. The modification of these parameters may be shown as: $\delta f_{1/2/3/3/4}^{'} = \delta f_{1/2/3/4} \pm (F_{av} \times \delta f_{1/2/3/4})$
3. Following the transfer learning, the modified delta function is being used in formant shifting [20] framework.

4. Result Evaluation

4.1. Experimental Setup

Three audio data sets are considered for testing and validating the proposed approach. These data sets span three different languages, namely: CHAINS (English), HINT (French) and VoxForge (German). Each data set comprises of 80 speech utterances of around 3 to 8 seconds. During cross-validation over extended data (LibreSpeech[2] for English, and VoxForge for German), some speech signals were deliberately taken longer (around 20 - 25 seconds) to account for variations in signal length. While evaluating the pitch, YAAPT [30] has been used, wherein the frame-size is kept at 25 milliseconds and the hop size is 50%. Further, four different noise environments have been considered for testing and validation of the proposed approach. These include Babble (multiple speech signals overlapping), Car (running at 50mph and 130mph) and Train (passing station at 50mph) [31]. These noises were added to both the original signal (O) and the modified signal (M) to get original noisy signal (O+N) and modified noisy signal (M+N). Thereafter, STOI evaluations were performed as follows:

$$O_{STOI} = STOI(O, O + N); \quad M_{STOI} = STOI(M, M + N)$$

Tables 2, 3 and 4 exhibit the major results for each target language data set, i.e., French, German and English respectively. Columns 3rd to 5th in these tables show the self improvement in STOI where the delta function generated through CLPSO was applied on the same dataset (e.g., in Table 2, $\mathrm{STOI_{M:M+N}}^{FRE \rightarrow FRE}$). Subsequently, the columns 6th to 7th and 8th to 9th exhibit the results when delta functions of different source datasets are generated through CLPSO, and applied on the target dataset after TL (e.g., in Table 2, $\mathrm{STOI_{M:M+N}}^{ENG \rightarrow FRE}$, and $\mathrm{STOI_{M:M+N}}^{GER \rightarrow FRE}$). It may be noted that all improvements are measured with respect to the case where no modifications were made to the formants (*NM*).

4.2. Transfer Learning from English to other Languages

From 6th and 7th columns of Tables 2 and 3, it can be seen that the transfer learning significantly improves the intelligibility for French and German languages respectively, in comparison to *NM*. These results are also compared against the self-improvement obtained through CLPSO based formant shifting. Further, the obtained results are consistent and clearly indicate an improvement in intelligibility across all noise environments and varying SNRs.

4.3. Transfer Learning from French and German

From 8th and 9th columns of Tables 2 and 3 respectively, and 6th to 9th columns of Table 4, an overall improvement in intelligibility was observed with respect to NM. However, in contrast to transfer learning from English as described in Section 4.2, we observed that in some cases, the modified delta function degraded the intelligibility instead of improving it. The plausible explanation could be attributed to the fundamental difference in the articulation of these languages. Most European languages (like French and German) are rather smooth in their vocation, i.e., the speech does not have large tonal variations. English, on the other hand is relatively rich in such variations. This might account for the fact that a model having parameters trained on CHAINS (English) data set performs well even when modified for other, rather smooth languages. The same phenomenon could also explain why the transfer of learning from smoother languages to tonally rich languages fails on account of not being trained for such variations to begin with. A more extensive study is deferred to future work to understand the underlying reasons for such a directional behaviour of transfer learning between languages.

4.4. Checking for Data Scarcity

To check for loopholes in the explanation given in Section 4.3, the German and English speech data sets were extended and the delta function parameters were re-estimated from these extended data sets. For this, 17 extra speech utterances of around 25 seconds each, were added to both German and English data sets. Due to space constraints, the validation is done for only Car_{50} noise condition, and the results of this experiment are presented in Table 5. These results demonstrate that increasing the data did not help and the decrease in STOI was not as a result of data scarcity.

Table 5: *STOI checked on extended English Data set (post transfer learning from extended German data set)*

Noise	SNR	$\mathrm{STOI_{O:O+N}}^{\mathrm{NM}}$	$\mathrm{STOI_{M:M+N}}^{GER \rightarrow ENG}$	Imp(%)
Car_{50}	-8	0.6519	0.5409	-17.0283%
	-14	0.5052	0.4420	-12.5100%
	-26	0.3236	0.3309	2.2613%

4.5. Cross dataset Validation

The results obtained from CHAINS data set was cross validated on another English data set (LibriSpeech). The results obtained seem to be along the lines of the results obtained from the CHAINS data set when the parameters were applied on itself (Table 6). Therefore, improvement in STOI when transfer learning is performed from English language is not dependent on the data set but on the language itself. Some demo examples and supplementary results may be found at the given link[3].

Table 6: *STOI checked on LibriSpeech (LIB) English Data set (post transfer learning from CHAINS (CHA))*

Noise	SNR	$\mathrm{STOI_{O:O+N}}^{\mathrm{NM}\,(LIB)}$	$\mathrm{STOI_{M:M+N}}^{CHA \rightarrow LIB}$	Imp(%)
Babble	-8	0.4751	0.5442	14.5389%
	-14	0.3756	0.4440	18.2093%
	-26	0.3077	0.3775	22.6620%
Car_{50}	-8	0.6836	0.7439	8.8297%
	-14	0.5484	0.6326	15.3604%
	-26	0.3656	0.4413	20.6963%
Car_{130}	-8	0.7510	0.7644	1.7870%
	-14	0.6403	0.6795	6.1114%
	-26	0.4474	0.5229	16.8889%
Train_{50}	-8	0.5611	0.6304	12.3574%
	-14	0.4643	0.5226	12.5728%
	-26	0.3509	0.4145	18.1338%

5. Conclusion

The proposed approach demonstrates improvement as high as 28.5% (French to German in presence of -26dB Babble noise). The highest improvement using CLPSO was 49.3% under same conditions. The transfer of parameters across languages is significantly faster than evaluating them using CLPSO. However, in a few cases, the STOI seems to decrease after Transfer Learning. Further improvement may be expected by using a smoother shaped delta function like Gaussian. The presented work has its applications in areas where improving intelligibility in somewhat real time across languages takes precedence over improving intelligibility maximally for individual languages.

6. Acknowledgement

This work is supported by SERB Govt. of India, through Project No. CRG/2018/003920 under the CRG-SERB Scheme.

[2] https://www.openslr.org/12

[3] https://tinyurl.com/44fad7cn

7. References

[1] K. Nathwani, F. Hafiz, A. Swain, R. Biswas, An optimal formant shifting approach for speech intelligibility enhancement, Submitted to: Computer Speech and Language, Elsevier (2020).

[2] R. Martin, Spectral subtraction based on minimum statistics, power 6 (8) (1994) 1182–1185.

[3] B. Xia, C. Bao, Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification, Speech Communication, Elsevier 60 (2014) 13–29. doi:https://doi.org/10.1016/j.specom.2014.02.001.

[4] T. Sreenivas, P. Kirnapure, Codebook constrained wiener filtering for speech enhancement, IEEE Transactions on Speech and Audio Processing 4 (5) (1996) 383–389. doi:10.1109/89.536932.

[5] Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing 32 (6) (1984) 1109–1121. doi:10.1109/TASSP.1984.1164453.

[6] Y. Jiang, H. Zhou, Z. Feng, Performance analysis of ideal binary masks in speech enhancement, in: 4th International Congress on Image and Signal Processing, Vol. 5, IEEE, 2011, pp. 2422–2425. doi:10.1109/CISP.2011.6100732.

[7] A. Zehtabian, H. Hassanpour, S. Zehtabian, V. Zarzoso, A novel speech enhancement approach based on singular value decomposition and genetic algorithm, in: IEEE International Conference of Soft Computing and Pattern Recognition, IEEE, 2010, pp. 430–435. doi:10.1109/SOCPAR.2010.5686627.

[8] P.-S. Huang, S. D. Chen, P. Smaragdis, M. Hasegawa-Johnson, Singing-voice separation from monaural recordings using robust principal component analysis, in: International Conference on Acoustics, Speech and Signal Processing, IEEE, 2012, pp. 57–60. doi:10.1109/ICASSP.2012.6287816.

[9] Y. Hu, P. Loizou, A subspace approach for enhancing speech corrupted by colored noise, IEEE Signal Processing Letters 9 (2002) 204–206. doi:10.1109/LSP.2002.801721.

[10] K. Nathwani, Joint acoustic echo and noise cancellation using spectral domain kalman filtering in double-talk scenario, in: 16th International Workshop on Acoustic Signal Enhancement (IWAENC), IEEE, 2018, pp. 1–330. doi:10.1109/IWAENC.2018.8521282.

[11] H. Padaki, K. Nathwani, R. M. Hegde, Single channel speech dereverberation using the lp residual cepstrum, in: National Conference on Communications (NCC), IEEE, 2013, pp. 1–5. doi:10.1109/NCC.2013.6487990.

[12] G. Kim, P. C. Loizou, Why do speech-enhancement algorithms not improve speech intelligibility?, in: International Conference on Acoustics, Speech and Signal Processing, IEEE, 2010, pp. 4738–4741. doi:10.1109/ICASSP.2010.5495169.

[13] K. Nathwani, G. Richard, B. David, P. Prablanc, V. Roussarie, Speech intelligibility improvement in car noise environment by voice transformation, Speech Communication, Elsevier 91 (2017) 17–27. doi:https://doi.org/10.1016/j.specom.2017.04.007.

[14] J. D. Griffiths, Optimum linear filter for speech transmission, The Journal of the Acoustical Society of America 43 (1) (1968) 81–86. doi:10.1121/1.1910768.

[15] R. J. Niederjohn, 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 24 (4) (1976) 277–282. doi:10.1109/TASSP.1976.1162824.

[16] C. Taal, J. Jensen, SII-based speech preprocessing for intelligibility improvement in noise, in: Annual Conference of the International Speech Communication Association, INTERSPEECH, 2013, pp. 3582–3586.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, A short-time objective intelligibility measure for time-frequency weighted noisy speech, in: International Conference on Acoustics, Speech and Signal Processing, IEEE, 2010, pp. 4214–4217. doi:10.1109/ICASSP.2010.5495701.

[18] C. Taal, R. Hendriks, H. Richard, Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure, Computer Speech & Language 28 (4) (2014) 858–872. doi:10.1016/j.csl.2013.11.003.

[19] R. Hendriks, J. Crespo, J. Jensen, C. Taal, Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time sii, Transactions on Audio, Speech, and Language Processing 23 (2015) 851–862. doi:10.1109/TASLP.2015.2409780.

[20] K. Nathwani, M. Daniel, G. Richard, B. David, V. Roussarie, Formant shifting for speech intelligibility improvement in car noise environment, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 5375–5379. doi:10.1109/ICASSP.2016.7472704.

[21] K. Nathwani, Intelligibility improvement using Kalman filtering & EM approach in formant shifting framework, in: International Symposium on Signal Processing and Information Technology (ISSPIT), IEEE, 2019, pp. 1–6. doi:10.1109/ISSPIT47144.2019.9001849.

[22] J. J. Liang, A. K. Qin, P. N. Suganthan, S. Baskar, Comprehensive learning particle swarm optimizer for global optimization of multimodal functions, Transactions on Evolutionary Computation 10 (3) (2006) 281–295. doi:10.1109/TEVC.2005.857610.

[23] M. Rahmati, R. Effatnejad, A. Safari, Comprehensive learning particle swarm optimization (clpso) for multi-objective optimal power flow, Indian Journal of Science and Technology 7 (3) (2014) 262–270. doi:10.17485/ijst/2014/v7i3.7.

[24] R. Patel, F. Hafiz, A. Swain, A. Ukil, Nonlinear excitation control of diesel generator: A command filter backstepping approach, Transactions on Industrial Informatics (2020) 1–1doi:10.1109/TII.2020.3017744.

[25] R. Biswas, K. Nathwani, F. Hafiz, A. Swain, An optimal near-end speech intelligibility improvement for variable car noise characteristics, Submitted to: Speech Communication, Elsevier (2020).

[26] R. Biswas, K. Nathwani, Optimal near-end speech intelligibility improvement using clpso based voice transformation in realistic noisy environments, Submitted to: Digital Signal Processing, Elsevier (2021).

[27] W. Shahzad, F. A. Khan, A. B. Siddiqui, Clustering in mobile ad hoc networks using comprehensive learning particle swarm optimization (clpso), in: International Conference on Future Generation Communication and Networking, Springer, 2009, pp. 342–349. doi:10.1007/978-3-642-10844-0_41.

[28] F. Cummins, M. Grimaldi, T. Leonard, J. Simko, The chains corpus: Characterizing individual speakers, in: SPECOM, Vol. 6, SPC RAS, 2006, pp. 431–435.

[29] M. Nilsson, S. D. Soli, J. A. Sullivan, Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise, The Journal of the Acoustical Society of America 95 (2) (1994) 1085–1099. doi:10.1121/1.408469.

[30] S. Zahorian, H. Hu, A spectral/temporal method for robust fundamental frequency tracking, The Journal of the Acoustical Society of America 123 (2008) 4559–71. doi:10.1121/1.2916590.

[31] D. Pearce, H.-G. Hirsch, The aurora experimental framework for the performance evaluations of speech recognition systems under noisy condition, in: Sixth International Conference on Spoken Language Processing, Vol. 4, INTERSPEECH, 2000, pp. 29–32.