Near-end Intelligibility Improvement Through Voice Transformation in Transfer Learning Framework

Ritujoy Biswas*, Karan Nathwani*, Vinayak Abrol‡
*Dept. of Electrical Engineering, IIT Jammu, India.
‡Infosys Centre for AI & Dept. of Computer Science and Engineering, IIIT Delhi, India.
Email: ritujoybiswas@gmail.com, karan.nathwani@iitjammu.ac.in, abrol@iiitd.ac.in

***Abstract*—In recent works, using voice transformation functions (VTF) in optimal shifting of formants has improved near-end speech intelligibility. Though these VTFs are promising, they are computationally expensive to optimize and generate unwanted artifacts during voice modification. Additionally, they were specific to the environmental condition they were optimized for. For the applicability of this approach to different languages without re-optimization, transfer learning (TL) was used to shape the parameters of VTF to accommodate the target language [1]. However, TL across noises and TL across languages and noises (simultaneously) was not viable due to the dependency on pitch information of source and target noises. Hence in this work, a statistical Gaussian Transformation Function (GTF) is developed with parameters optimized for specific environmental conditions. Defined by just three parameters, the optimization time came down, and the intelligibility surpassed the previously used VTF. Additionally, GTF allows TL across both noises and languages simultaneously, with fewer artifacts while shifting the formants.**

***Index Terms*—Speech Intelligibility, Gaussian Transformation Function, CLPSO, STOI, Transfer Learning**

I. Introduction

The objective of any speech communication system (viz., hearing aids, mobile telephony, and public address system) is to improve the intelligibility of spoken words in noisy environments. Consider a case where noise is present in both far-end (talker) and near-end environments (listener). The impact of far-end noise can typically be eliminated by mono-aural speech enhancement techniques [2], [3] (not the focus of this work). In this work, we pre-process the far-end speech (assuming successful noise suppression in far-end signal) before being played back in near-end noise [4]–[6].

Most works on near-end speech intelligibility deal with processing speech characteristics like pitch, formants, average power, and voiced/unvoiced segments [7]–[9] in a single microphone setting. However, some recent works are governed by improving the articulation and co-articulation of sound features like plosive, vowels, consonants [10], [11], and phonology [12]. The works in [4], [5] maximize the speech intelligibility index (SII) by redistributing the energy according to the perceptual distortion measure and by designing the optimal linear filter, respectively. The authors in [13] proposed a blind acoustic mask that identified and selected the speech samples with lower noise-to-speech proportion by deriving adaptive information using noise statistics. The aforementioned methods assume that the noise statistics are known beforehand and perform poorly at low SNRs.

To address this, [14] optimized the shaping parameters of the voice transformation function (VTF) using Comprehensive Learning Particle Swarm Optimization (CLPSO), which optimizes the shift in formants to improve intelligibility. However, it generates significant artifacts due to aggressive formant shifts in the unvoiced frames [8], which degrades intelligibility. The changes in languages or noises were also not addressed, requiring computationally intensive re-optimizing of the parameters for the new conditions. Hence, [1] proposed rapidly modifying the parameters optimized for a particular language and transferring to a different one. The authors exploited the comparative pitch and formant information of the target language and the one on which the shaping parameters were initially optimized (i.e., source). However, transfer across noises was not possible, and this transfer across languages favored English as a source language. This directional nature may be attributed to the differences in phonetic organization and speech production across languages [15]–[17].

Thus, the major contributions of this work answer the following questions: a) Is there any voice transformation function that reduces the optimization time of CLPSO while generating fewer artifacts (Section II)? b) Is it possible to transfer across noises using some statistically guided transformation function (Section III)? c) How to achieve transfer across both languages and noises simultaneously (Section III)? The results and conclusions are mentioned in Sections V & VI, respectively.

II. CLPSO-based Intelligibility Improvement using Gaussian Transformation Function (GTF)

This section highlights our first contribution as the development of a statistical Gaussian transformation function (GTF). The shaping parameters (as shown in Figure 1) are evaluated using a variant of particle swarm optimization known as CLPSO. Although, CLPSO has already been used in our previous work (see [14], [18]) to optimize the VTF of trapezoidal shape (termed as TTF) having five parameters. In contrast, the GTF is defined by its three parameters - μ, σ & h. The mean (μ) and standard deviation (σ) decide the positioning

This work is supported by IIITD-IITD joint research grant (MFIRP-233) and Infosys Foundation via Infosys Centre for AI, IIIT Delhi.

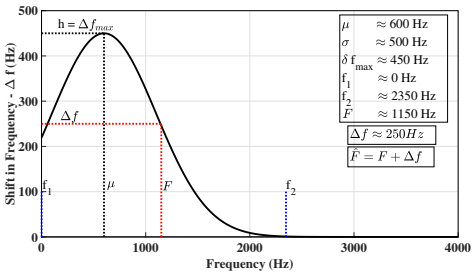


Fig. 1. A typical GTF illustrating formant shifting.

and spread of the TF over the spectrum. The maximum shift in a formant (Δf_{max}) is given by the height (h) of the GTF.

CLPSO [14] starts with a random set of parameters used to perform Formant Shifting (FS). The resulting modified speech is compared with the unmodified speech in terms of STOI. This comparison gives a sense of '*fitness*', which assigns a penalty to the current parameter set. This decides the direction of movement of the search algorithm to find the best parameters. This procedure is repeated iteratively, and the parameters get regularly updated. The resulting shaping parameters obtained under different environmental conditions are given in Table 1. After optimization, the shift in each formant is controlled by these three 'shaping' parameters as:

$$\hat{F} = \begin{cases} F + \{\Delta f = \mathcal{G}(F \mid \mu, \sigma, h)\}, & if \quad f_1 \leq F \leq f_2 \\ F + \{\Delta f = 0\}, & otherwise \end{cases} \tag{1}$$

where:

$$\begin{aligned} f_1 &= \mathbf{max}(0\ Hz\ ,\ [(f_H\ <\ \mu)\ where\ h = 0]) \\ f_2 &= \mathbf{min}([(f_L\ >\ \mu)\ where\ h = 0]\ ,\ 4000\ Hz) \end{aligned} \tag{2}$$

'f_1' & 'f_2' decide the effective range of the GTF. The value of f_1 is decided such that it is either 0 Hz or the highest frequency (f_H) below the mean (μ) where the effective GTF starts - whichever is the higher frequency. The value of f_2 is similarly decided, such that it is either 4000 Hz or the lowest frequency (f_L) above the mean (μ) where the effective GTF ends - whichever is lower. 'F' & '$\hat{F}$' are the original and shifted formants, respectively. $\mathcal{G}(F \mid \mu, \sigma, h)$ refers to the shift

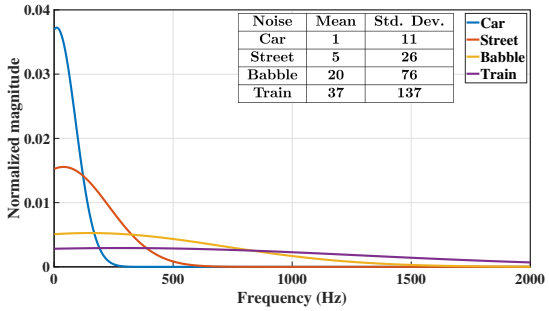


Fig. 2. Gaussian approximations of noise spectra with statistics.

TABLE I
GTF PARAMETERS GENERATED VIA CLPSO.

Noise	SNR (dB)	English (*EN*) Mean (μ)	English (*EN*) Height (h)	English (*EN*) Std. Dev (σ)	French (*FR*) Mean (μ)	French (*FR*) Height (h)	French (*FR*) Std. Dev (σ)	German (*GE*) Mean (μ)	German (*GE*) Height (h)	German (*GE*) Std. Dev (σ)
Babble (*BB*)	-8	970.92	487.93	538.33	1421.42	469.44	994.77	294.44	479.11	109.42
	-14	1075.43	498.74	661.81	951.78	430.42	998.17	302.84	487.82	122.29
	-26	1285.88	498.64	734.02	777.78	461.19	849.67	264.46	491.17	92.61
Car_{130} (*CR*)	-8	984.38	497.00	373.09	879.43	439.18	355.10	461.50	473.07	139.59
	-14	890.95	493.90	402.73	1328.49	493.02	645.31	460.73	498.98	171.28
	-26	1079.43	496.35	682.76	1108.12	436.89	987.59	344.75	498.32	150.36
Street (*ST*)	-8	990.66	474.88	527.54	1416.10	498.14	706.18	355.15	489.13	136.22
	-14	1007.48	490.52	570.30	1142.14	439.71	983.26	336.34	491.17	137.12
	-26	1204.48	499.06	686.39	455.38	481.99	940.51	295.91	488.84	116.17
$Train_{50}$ (*TR*)	-8	1018.01	487.78	565.87	837.00	413.46	739.82	293.44	468.67	113.40
	-14	1213.24	497.22	674.49	959.28	426.09	956.59	269.10	498.05	92.65
	-26	1331.26	498.83	741.32	565.29	499.31	972.54	314.79	496.90	126.24

(Δf) in formant at the formant location F, when the GTF is defined by the parameters $\mu, \sigma,\ \&\ h(\Delta f_{max})$ (Figure 1).

The shifted formants ($\hat{F}$) are then used to reconstruct the modified signal ($\hat{s}(n)$) as given in [14]. After energy normalization of the modified signal with respect to the original signal ($s(n)$) and the noise ($u(n)$), the normalized modified signal ($\bar{s}(n)$) and the original signal ($s(n)$) are both added with noise to generate $\bar{s}_u(n)$ and $s_u(n)$. STOI is calculated between $\{s_u(n)\ \&\ s(n)\}$ and between $\{\bar{s}_u(n)\ \&\ \bar{s}(n)\}$. If the STOI for modified speech is more than the unmodified speech for the same environmental conditions, it indicates an improvement in intelligibility.

The optimal shaping parameters are given in Table I. These parameters of GTF have been evaluated through CLPSO for all combinations of languages: (English (*EN*) [19], French (*FR*) [20] & German (*GE*) (VoxForge[1]), noise types: Babble (*BB*), Car (*CR*), Street (*ST*), & Train (*TR*) from NOIZEUS database [21]) and SNR levels (-8dB, -14dB, & -26dB). The variation in shaping parameters across languages indicates a difference in the production of utterances across languages [15]–[17].

III. Transfer Learning via GTF

As mentioned in Section II, the optimization of GTF parameters is specific to a fixed language, noise type, and SNR level. Even with reduced complexity, it is impractical to re-optimize for every new combination. To that end, the following contributions are highlighted as follows: (a) TL across noises using GTF, (b) combining TL across languages using GTF (earlier achieved through a trapezoidal transformation function (TTF) [1]) with TL across noises. Notably, while optimizing the GTF parameters with CLPSO takes hours, the modification through TL only takes a few minutes.

A. Transfer learning across noises

In this work, four noises of varying stationarity are considered (Figure 2): Babble (*BB*), a car engine (*CR*), a busy street (*ST*), and a train entering a station (*TR*) [21]. The TL using TTF required formant and pitch information to compare source and target environments. Since these realistic noises have no pitch, to achieve TL across noises, some other statistical comparative criteria among noises were required, like mean frequency (μ_n) and spectral spread (σ_n). A Gaussian distribution was fit over the magnitude spectrum of each noise, and their statistics (μ_n and σ_n) were normalized with

[1]http://www.voxforge.org/home

respect to *CR* (lowest values of statistics). Thereafter, the GTF generated using CLPSO for a source noise is modified for different target noises using these comparative statistics as modification factors (keeping language and SNR fixed).

If the modification factor: $\frac{(\mu_{tn}-\mu_{sn})}{\mu_{sn}} \geq 1$, the GTF generated for the *source* noise is shifted to the right, as the target noise is more prevalent higher up the spectrum. Similarly, if the factor < 1, the GTF is shifted to the left. If the modification factor: $\frac{(\sigma_{tn}-\sigma_{sn})}{\sigma_{sn}} \geq 1$, the spread of GTF generated for the *source* noise is increased, as the target noise is spread over a larger region of the spectrum. Similarly, if the factor < 1, the standard deviation of the GTF is reduced. These operations can be represented as:

$$\begin{aligned} \mu_T &= \mu_S + \frac{\mu_{tn} - \mu_{sn}}{\mu_{sn}}\% \; of \; \mu_S \\ \sigma_T &= \sigma_S + \frac{\sigma_{tn} - \sigma_{sn}}{\sigma_{sn}}\% \; of \; \sigma_S \end{aligned} \tag{3}$$

Here, μ_S (μ_T) and σ_S (σ_T) denote the mean and standard deviation of the GTF in the source (target) noise environment, respectively. The terms μ_{sn} (μ_{tn}) and σ_{sn} (σ_{tn}) denote the mean and standard deviation of the Gaussian approximation of the source (target) noise magnitude spectrum, respectively. The deviation in GTF after TL from direct optimization (CLPSO) is illustrated in Figure 3. After TL, formant shifting is re-employed via GTF to obtain the modified signal.

B. Transfer learning across languages & noises

Another contribution of this paper combines the TL across languages with TL across noises using GTF, which was not possible using TTF. This is detailed in Algorithm 1.

$F_{S_{avg}}$ and $F_{T_{avg}}$ denote the average formant values of speech in the *source* and *target* languages. The mean of the GTF in the target noise environment (μ_T) is handled by transfer across languages and noises and is decided by the minimum of two modifications factors: $\mathcal{A}$ and $\mathcal{B}$, as given in Algorithm 1. However, the standard deviation of the GTF in the target noise environment (σ_T) is determined solely by transfer across noise and is decided by the comparison of the standard deviations of the Gaussian approximations of the noise spectral distributions (σ_{sn} & σ_{tn}). After generating the GTF via TL, formant shifting is re-employed to obtain the modified signal.

Algorithm 1: TL across languages and noises

Input : $\mu_S, F_{T_{avg}}, F_{S_{avg}}, \mu_{tn}, \mu_{sn}, \sigma_{tn}, \sigma_{sn}$
Output: μ_T, σ_T

Evaluate Modification Factor $\mathcal{A}$: $\frac{F_{T_{avg}}}{F_{S_{avg}}} \times \mu_S$
Evaluate Modification Factor $\mathcal{B}$: $\frac{(\mu_{tn}-\mu_{sn})}{\mu_{sn}}$ % of μ_S
if *sgn*($\mathcal{A}$) == *sgn*($\mathcal{B}$) **then** /* check sign */
    $\mu_T = \mu_S + \min(\mathcal{A}, \mathcal{B})$
else
    if $|\mathcal{B}| < |\mathcal{A}|$ **then** /* compare magnitude */
        $\mu_T = \mu_S + \mathcal{B}$
    else
        $\mu_T = \mu_S + \mathcal{A}$
    end
end
Evaluate Modification Factor $\mathcal{C}$: $\frac{(\sigma_{tn}-\sigma_{sn})}{\sigma_{sn}}$ % of σ_S
$\sigma_T = \sigma_S + \mathcal{C}$

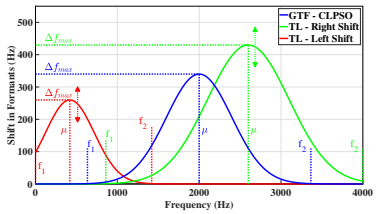


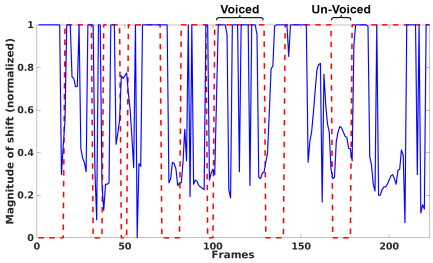
Fig. 3. GTF generated using CLPSO and TL after Left and right shift.

IV. Experimental Setup and Results

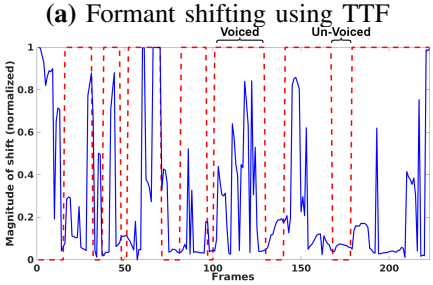
Experiments were performed using EN, FR, and GE languages, in BB, CR, ST, and TR noises at -8, -14, and -26 dB SNR levels. For each language, 80 speech sequences of 2 to 5 seconds were used, and all results were averaged for the 80 sequences. For formant extraction, the speech is first divided into short-time frames through a Hanning window of 25 ms duration with 50% overlap with the subsequent frame. Pitch is evaluated for the frames using the algorithm mentioned in [22]. In the tables, *O+N* (*M+N*) denotes original (modified) speech in the presence of noise.

A. Significance of GTF over TTF in CLPSO & TL

The transition from TTF to GTF reduced the number of parameters to be optimized and decreased the optimization time. Earlier, CLPSO converged in about 68 hours to optimize 5 shaping parameters of TTF. Now, it converged in about 25 hours for 3 shaping parameters of GTF. The comparative study of FS through TTF and GTF reveals an interesting observation. Figure 4 shows the shifting of formants using TTF to be much



(a) Formant shifting using TTF



(b) Formant shifting using GTF

Fig. 4. Comparative shift in formants on an English sentence.

TABLE II
MEAN STOI CHECKED ON *BB* (SELF, POST TRANSFER LEARNING FROM *CR*, *ST* & *TR*). LANGUAGE - *EN*.

SNR (dB)	Self (CLPSO via GTF on *BB*)		CLPSO on *CR*, *ST* & *TR* and transfer to *BB*		
	$STOI_{O+N}^{BB}$	$STOI_{M+N}^{BB\rightarrow BB}$	$STOI_{M+N}^{CR\rightarrow BB}$	$STOI_{M+N}^{ST\rightarrow BB}$	$STOI_{M+N}^{TR\rightarrow BB}$
-8	0.41	0.55 (+32.27%)	0.48 (+16.16%)	0.54 (+31.27%)	**0.55 (+32.26%)**
-14	0.31	0.44 (+41.32%)	0.42 (+33.03%)	**0.44 (+40.79%)**	0.44 (+40.61%)
-26	0.25	0.35 (+36.50%)	0.34 (+34.39%)	**0.35 (+36.41%)**	0.35 (+36.22%)

TABLE III
MEAN STOI CHECKED ON *ST* (SELF, POST TRANSFER LEARNING FROM *BB*, *CR* & *TR*). LANGUAGE - *EN*.

SNR (dB)	Self (CLPSO via GTF on *ST*)		CLPSO on *BB*, *CR* & *TR* and transfer to *ST*		
	$STOI_{O+N}^{ST}$	$STOI_{M+N}^{ST\rightarrow ST}$	$STOI_{M+N}^{BB\rightarrow ST}$	$STOI_{M+N}^{CR\rightarrow ST}$	$STOI_{M+N}^{TR\rightarrow ST}$
-8	0.52	0.63 (+21.88%)	0.63 (+21.14%)	0.57 (+9.57%)	**0.63 (+21.30%)**
-14	0.40	0.53 (+32.65%)	0.53 (+32.09%)	0.49 (+22.64%)	**0.53 (+32.40%)**
-26	0.31	0.41 (+32.77%)	**0.41 (+32.72%)**	0.41 (+32.59%)	0.41 (+32.53%)

more aggressive (especially for unvoiced frames) than using GTF for the same audio. The high (low) pulses in the dotted red line indicate voiced (unvoiced) frames in Figure 4. An aggressive formant shifting in unvoiced frames is known to be counterproductive for intelligibility improvement [8]. For statistical verification, the mean and the standard deviation of the difference in normalized formant shifts between TTF and GTF ($FS_{TTF} - FS_{GTF}$) were evaluated for the unvoiced frames of all audio files across all combinations of the aforementioned languages, noises, and SNR levels. These mean values lie in the range [0.02:0.41], while the standard deviation lies in the range [0.02:0.12]. The lesser shift in case of FS_{GTF} relative to FS_{TTF} indicates the consistently lesser modification in unvoiced frames using GTF, thereby inducing much fewer artifacts than TTF.

B. Performance of TL across Noises

The results of TL across noises (keeping language fixed) are given in Tables II and III, where TL is applied to *BB* and *ST* noises, respectively. For instance, in Table II, GTF is optimized individually for *CR*, *ST*, and *TR* using CLPSO, and TL is applied from each of these noises to *BB*. In general, it was observed that learning from non-stationary noises performed better than from relatively more stationary ones. Figure 2 shows the decreasing stationarity of noises from *CR* to *TR*.

This trend, however, discontinues as SNR decreases. In Tables II and III, it can be seen that the improvements in intelligibility to the target noise are nearly identical irrespective of the source noise when the SNR is -26 dB. This indicates that at very low SNRs, the proposed approach is independent of the source noise type. It is also evident that the GTF, when used in the TL framework, works reasonably well, close to the performance of direct optimization by CLPSO. The values in brackets represent the percentage improvement in STOI values with respect to the no modification case (column 2 of II and III). The bold values in all tables represent the highest improvement in mean STOI value at that particular SNR.

C. Performance of TL across languages and noises

To illustrate the performance of TL across languages and noises, we consider 4 cases, as shown in Tables IV to VII. For instance, in Table IV, the case *EN*(*TR*)→*FR*(*BB*) indicates that the CLPSO via GTF is optimized for English (train) and

TABLE IV
TL FOR *EN*(*TR*)→*FR*(*BB*) AND COMPARISON WITH [1] & [14].

SNR (dB)	*FR(BB)* $STOI_{O+N}^{NM}$	CLPSO (TTF) $STOI_{M+N}^{FR(BB)}$	CLPSO (GTF) $STOI_{M+N}^{FR(BB)}$	TL-Lang (TTF) $STOI_{M+N}^{EN(BB)\rightarrow FR(BB)}$	TL (Lang + Noise) (GTF) $STOI_{M+N}^{EN(TR)\rightarrow FR(BB)}$
-8	0.44	0.57 (+29.32%)	**0.58 (+31.32%)**	0.47 (+6.68%)	**0.56 (+27.27%)**
-14	0.31	0.45 (+42.67%)	**0.47 (+51.61%)**	0.34 (+9.42%)	**0.44 (+41.94%)**
-26	0.25	0.34 (+39.57%)	**0.35 (+40.00%)**	0.27 (+8.00%)	**0.32 (+28.00%)**

TABLE V
TL FOR *EN*(*TR*)→*FR*(*CR*) AND COMPARISON WITH [1] & [14].

SNR (dB)	*FR(CR)* $STOI_{O+N}^{NM}$	CLPSO (TTF) $STOI_{M+N}^{FR(CR)}$	CLPSO (GTF) $STOI_{M+N}^{FR(CR)}$	TL-Lang (TTF) $STOI_{M+N}^{EN(CR)\rightarrow FR(CR)}$	TL (Lang + Noise) (GTF) $STOI_{M+N}^{EN(TR)\rightarrow FR(CR)}$
-8	0.76	0.79 (+4.80%)	0.79 **(+4.87%)**	0.77 (+0.19%)	**0.79 (+3.94%)**
-14	0.62	0.70 (+12.99%)	**0.70 (+13.57%)**	0.63 (+2.05%)	**0.70 (+12.90%)**
-26	0.39	0.53 (+36.74%)	**0.54 (+38.85%)**	0.42 (+7.85%)	**0.53 (+35.89%)**

TABLE VI
TL FOR *GE*(*TR*)→*EN*(*BB*) AND COMPARISON WITH [1] & [14].

SNR (dB)	*EN(BB)* $STOI_{O+N}^{NM}$	CLPSO (TTF) $STOI_{M+N}^{EN(BB)}$	CLPSO (GTF) $STOI_{M+N}^{EN(BB)}$	TL-Lang (TTF) $STOI_{M+N}^{GE(BB)\rightarrow EN(BB)}$	TL (Lang + Noise) (GTF) $STOI_{M+N}^{GE(TR)\rightarrow EN(BB)}$
-8	0.41	**0.55 (+32.30%)**	0.55 (+32.27%)	0.39 (-5.27%)	**0.51 (+22.20%)**
-14	0.31	**0.45 (+42.73%)**	0.44 (+41.32%)	0.33 (+4.12%)	**0.38 (+22.54%)**
-26	0.25	**0.36 (+42.31%)**	0.35 (+36.50%)	0.27 (+5.75%)	**0.29 (+15.82%)**

TABLE VII
TL FOR *FR*(*TR*)→*EN*(*CR*) AND COMPARISON WITH [1] & [14].

SNR (dB)	*EN(CR)* $STOI_{O+N}^{NM}$	CLPSO (TTF) $STOI_{M+N}^{EN(CR)}$	CLPSO (GTF) $STOI_{M+N}^{EN(CR)}$	TL-Lang (TTF) $STOI_{M+N}^{FR(CR)\rightarrow EN(CR)}$	TL (Lang + Noise) (GTF) $STOI_{M+N}^{FR(TR)\rightarrow EN(CR)}$
-8	0.73	0.76 (+4.11%)	**0.76 (+4.72%)**	0.62 (-15.07%)	**0.70 (-4.11%)**
-14	0.61	0.68 (+11.48%)	**0.68 (+12.69%)**	0.54 (-11.48%)	**0.61 (0%)**
-26	0.39	0.52 (+33.33%)	**0.53 (+35.18%)**	0.43 (+10.26%)	**0.46 (+17.95%)**

TL is applied to French(babble). Since *TR* and *BB* noises have similar spectra (refer Figure 2), TL from *TR* to *CR* is considered in Table V and VII to remove ambiguity. Column 2 in Tables IV to VII indicate the STOI of no modification case. Column 3 in these Tables indicates the results of TTF optimized using CLPSO [14], and column 4 indicates the results of GTF optimized using CLPSO. Column 5 shows the results of TL across languages using TTF [1], and column 6 indicates the results of TL across languages and noises using GTF. It can be seen that similar to TL across noises, TL across languages and noises produces results close to the performance of direct optimization by CLPSO. Throughout Tables IV to VII, it is a noteworthy observation that while optimization of GTF using CLPSO leads to minor improvements in intelligibility over TTF, the performance boost achieved by TL across languages and noises is significantly more than TL across languages alone. Due to rounding off, some % changes might differ though the difference in values appears the same.

D. Directional in-dependence during TL across languages

In [1], TL across languages exhibited a directional nature in terms of preference for the source language. The performance of TL using English as a source language was better than in other cases. When French or German were used as source languages, the performance dropped significantly, often degrading intelligibility instead of improving it. This behavior may be due to varying degrees of phonetic richness and differences in speech production across languages [15]–[17].

However, simultaneous TL across languages and noises compensates for this behavior. This is probably because acclimatizing to changes in noise results in a higher degree of intelligibility improvement than adapting to changes in language. It may also be attributed to the statistical GTF used in the TL

TABLE VIII
MEAN OPINION SCORES FOR *GE(TR)→EN(BB)*.

SNR	MOS_{O}^{EN} NM	$MOS_{(O+N)}^{EN(BB)}$ NM	$MOS_{(M+N)}^{EN(BB)}$ CLPSO (TTF)	$MOS_{(M+N)}^{EN(BB)}$ CLPSO (GTF)	$MOS_{(M+N)}^{GE(TR)\rightarrow EN(BB)}$ TL (Lang + Noise)
-8	5	2.1000	2.4739	**3.4594**	3.3083
-14	5	1.7782	2.1565	2.2042	**2.3478**
-26	5	1.5304	**1.9217**	1.6875	1.5739

framework. This is evident through the analysis presented in Tables VI and VII. It can also be seen that in most cases, the performance that was degrading in TL across languages, improved markedly through TL across languages and noises. For additional results, see https://github.com/Ritujoy/UTL_results.

E. Mean opinion scores (MOS) evaluation

We collected the opinion of 23 people aged 18 to 27 years on cases of TL, where the target language was English, as most listeners were proficient only in English. Results of the case: *GE(TR)→EN(BB)* are given in Table VIII. The listeners graded the intelligibility of the audio files on a scale of 1 to 5 (1 being negligibly intelligible and 5 being fully intelligible). First, 10 clean audio files were played for the listeners to set the reference (MOS_{O}^{EN}). Thereafter, they were given the same signals mixed with noise ($MOS_{(O+N)}^{EN(BB)}$). Next, their opinions were recorded for the case where the signals were modified through direct optimization using CLPSO (TTF & GTF) and mixed with noise ($MOS_{(M+N)}^{EN(BB)}$). Finally, their opinions were recorded for the signals modified after TL, and mixed with noise ($MOS_{(M+N)}^{GE(TR)\rightarrow EN(BB)}$). It can be seen that the MOS values through TL across languages and noises are close to the direct optimization using CLPSO (GTF). As SNR increases, the artifacts become more prominent, especially for TTF than in GTF. These MOS values are corroborated by Table VI.

V. CONCLUSIONS

The proposed work improves near-end speech intelligibility in varying languages, noises, and SNRs. The Gaussian Transformation Function (GTF), which replaced the trapezoidal transformation function (TTF) in the formant shifting (FS), reduced the time complexity due to the optimization of fewer parameters. The performance of GTF surpassed TTF and generated fewer artifacts through restrained modifications in unvoiced frames. Transfer learning (TL) across noises was made possible due to the statistical shape (Gaussian) of the voice transformation. Finally, through transfer across languages and noises simultaneously, intelligibility was extensively improved, while mitigating the directional nature of TL across languages. The proposed work has immense applications, especially in cases where rapid intelligibility enhancement across varying conditions is more critical than maximizing intelligibility for one fixed environment.

REFERENCES

[1] R. Biswas, K. Nathwani, and V. Abrol, "Transfer Learning for Speech Intelligibility Improvement in Noisy Environments," in *INTERSPEECH*, pp. 176–180, ISCA, 2021.

[2] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 5059–5063, IEEE, 2018.

[3] Q. Liu, W. Wang, P. J. Jackson, and Y. Tang, "A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions," in *European Signal Processing Conference (EUSIPCO)*, pp. 1270–1274, IEEE, 2017.

[4] C. Taal and J. Jensen, "SII-based speech preprocessing for intelligibility improvement in noise," in *INTERSPEECH*, pp. 3582–3586, ISCA, 2013.

[5] C. Taal, R. Hendriks, and H. Richard, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Computer Speech & Language*, vol. 28, no. 4, pp. 858–872, 2014.

[6] A. J. Fuglsig, J. Østergaard, J. Jensen, L. S. Bertelsen, P. Mariager, and Z.-H. Tan, "Joint far-and near-end speech intelligibility enhancement based on the approximated speech intelligibility index," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 7752–7756, IEEE, 2022.

[7] S. Shobha and R. Rajavel, "Improving speech intelligibility in monaural segregation system by fusing voiced and unvoiced speech segments," *Circuits, Systems, and Signal Processing*, vol. 38, no. 8, pp. 3573–3590, 2019.

[8] K. Nathwani, G. Richard, B. David, P. Prablanc, and V. Roussarie, "Speech intelligibility improvement in car noise environment by voice transformation," *Speech Communication*, vol. 91, pp. 17–27, 2017.

[9] M. Song, F. Chen, X. Wu, and J. Chen, "A time-weighted method for predicting the intelligibility of speech in the presence of interfering sounds," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 5589–5593, IEEE, 2018.

[10] E. Koffi, *Relevant acoustic phonetics of L2 English: Focus on intelligibility*. CRC Press, 2021.

[11] J. Fritsch and M. Magimai-Doss, "Utterance verification-based dysarthric speech intelligibility assessment using phonetic posterior features," *Signal Processing Letters*, vol. 28, pp. 224–228, 2021.

[12] J. S. Almeida, P. P. Rebouças Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliūnas, and V. H. C. de Albuquerque, "Detecting parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognition Letters*, vol. 125, pp. 55–62, 2019.

[13] F. Farias and R. Coelho, "Blind adaptive mask to improve intelligibility of non-stationary noisy speech," *IEEE Signal Processing Letters*, vol. 28, pp. 1170–1174, 2021.

[14] R. Biswas and K. Nathwani, "Optimal near-end speech intelligibility improvement using CLPSO-based voice transformation in realistic noisy environments," *Circuits, Systems and Signal Processing*, vol. 41, p. 6999–7034, 2022.

[15] A. Valdman and P. Delattre, "Comparing the phonetic features of english, french, german and spanish," *The Modern Language Journal*, vol. 51, no. 7, pp. 430–431, 1965.

[16] W. Strange, A. Weber, E. S. Levy, V. Shafiro, M. Hisagi, and K. Nishi, "Acoustic variability within and across german, french, and american english vowels: Phonetic context effects," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1111–1129, 2007.

[17] J. H. Hansen, M. Bokshi, and S. Khorram, "Speech variability: A cross-language study on acoustic variations of speaking versus untrained singing," *The Journal of the Acoustical Society of America*, vol. 148, no. 2, pp. 829–844, 2020.

[18] K. Nathwani, F. Hafiz, A. Swain, and R. Biswas, "Speech intelligibility enhancement using an optimal formant shifting approach," in *International Symposium on Image and Signal Processing and Analysis*, pp. 120–125, IEEE, 2021.

[19] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains corpus: CHaracterizing INdividual Speakers," in *The International Conference on Speech and Computer*, vol. 6, pp. 431–435, SPC RAS, 2006.

[20] M. Nilsson, S. D. Soli, and 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*, vol. 95, no. 2, pp. 1085–1099, 1994.

[21] H. Yi and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588–601, 2007.

[22] S. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, vol. 123, pp. 4559–4571, 2008.