Bypassing Optimization Complexity through Transfer Learning & Deep Neural Nets for Speech Intelligibility Improvement

Ritujoy Biswas
Indian Institute of Technology Jammu
ritujoybiswas@gmail.com

Abstract

This extended abstract highlights the research ventures and findings in the domain of speech intelligibility improvement. Till this point, an effort has been to simulate the Lombard effect, which is the deliberate human attempt to make a speech more intelligible when speaking in the presence of interfering background noise. To that end, an attempt has been made to shift the formants away from the noisy regions in spectrum both sub-optimally and optimally. The sub-optimal shifting methods were based upon Kalman filtering and EM approach. The optimal shifting involved the use of optimization to maximize an objective intelligibility index after shifting the formants. A transfer learning framework was also set up to bring down the computational complexity.

1 Motivation of Research

While much of the research focus has been on improvement in quality of speech signals, certain applications call for proper intelligibility of the speech rather than how pleasing it is to the listener. Thus, the prime motivation of the current research is to ensure that information is not lost to noise and is communicated in a robust manner, especially at very low SNR levels.

2 Key Issues; Identified and Addressed

The most common causative factor of loss in speech intelligibility is the presence of background noise. When noise occupies the same regions of the spectrum as speech, the intelligibility falls drastically. One solution is to shift the formants away from the noisy regions in the spectrum. The result would be that the information content in those formants would also be shifted away from the noise and would thus cease to be afflicted by it. The details of formant shifting are given in (Nathwani et al., 2016).

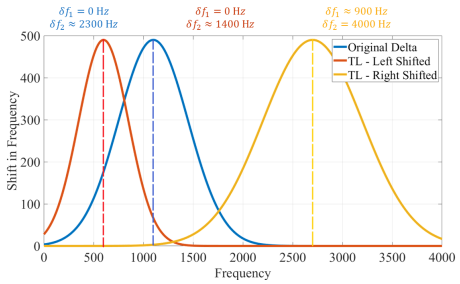

Figure 1: Transformation Function (TF) obtained through CLPSO and Transfer Learning (TL) via left (L) and right (R) shifting.

3 Major Contributions

Following have been the major contributions so far:

1. The first major contribution was the implementation of Comprehensive Learning Particle Swarm Optimization (CLPSO) (Rahmati et al., 2014) for optimization of 5 parameters a Trapezoidal Delta Function based formant shifting framework.

2. One drawback of this optimization was the enormous time-complexity which rendered the approach unsuitable for real-time applications. To address this issue, the next contribution was made. A Transfer Learning (TL) framework was developed which transferred the learning across languages.

3. In parallel, another attempt was made to reduce the time complexity by replacing the Trapezoidal formant shifting with a Gaussian one. Therefore, the next contribution was training a Gaussian delta function using CLPSO to optimize a set of 3 (instead of 5) shaping parameters.

4. Since Gaussian is a statistical shape, it allowed the incorporation of noise in the TL

Proceedings of the 18th International Conference on Natural Language Processing, pages 651–653
Silchar, India. December 16 - 19, 2021.

Table 1: Universal TL performance measured on STOI for EN(TR)→FR(BB) and compared with Trap based TL

SNR	$STOI_O^{NM}$ FR(*BB*)	CLPSO (Trap) $STOI_M^{FR(BB)\rightarrow FR(BB)}$	TL-Lang (Trap) $STOI_M^{EN(BB)\rightarrow FR(BB)}$	CLPSO (Gauss) $STOI_M^{FR(BB)\rightarrow FR(BB)}$	TL-Noise (Gauss) $STOI_M^{FR(TR)\rightarrow FR(BB)}$	TL-Univ (Gauss) $STOI_M^{EN(TR)\rightarrow FR(BB)}$
-8	0.44	0.57 (+29.54%)	0.47 (+6.82%)	0.58 (+31.82%)	0.57 (+29.54%)	0.56 (+27.27%)
-14	0.31	0.45 (+45.16%)	0.34 (+9.68%)	0.47 (+51.61%)	0.47 (+51.61%)	0.44 (+41.94%)
-26	0.25	0.34 (+36.00%)	0.27 (+8.00%)	0.36 (+44.00%)	0.35 (40.00%)	0.32 (+28.00%)

framework which was earlier not considered. Thus, the next contribution was the development of another TL framework for transferring the learning from one noise environment (*source*) to another (*target*).

5. The final contribution was the amalgamation of both the TL approaches to form a universal TL framework.

4 Methodologies

The foundation of the current work lies in shifting the formants away from noisy regions in speech. To that end, the mathematical representation of this shifting through a Gaussian delta function can be represented as shown:

$$\hat{F} = \begin{cases} \frac{h}{(\mu - \delta f_1)} + F, & if \quad \delta f_1 \leq F < \mu \\ \frac{-h}{(\delta f_2 - \mu)} + F, & if \quad \mu \leq F \leq \delta f_2 \\ F, & otherwise \end{cases}$$
$$\delta f_1 = \max(0\ Hz,\ f(< \mu\ @\ TF = 0))$$
$$\delta f_2 = \min(f(> \mu\ @\ TF = 0),\ 4000\ Hz) \quad (1)$$

For the universal TL framework, the formant shifting is applied after the parameters have been modified for the new combination of language and noise at a certain SNR. This universal TL can be mathematically represented as:

$$\mu_T = \mu_S \pm \min(|\frac{F_{T_{avg}}}{F_{S_{avg}}} \times \mu_S|, |\frac{\mu_{tn} \sim \mu_{sn}}{\mu_{sn}}\% \text{ of } \mu_S|) \quad (2)$$
$$\sigma_T = \sigma_S \pm \frac{\sigma_{tn} \sim \sigma_{sn}}{\sigma_{sn}}\% \text{ of } \sigma_S$$

The subscripts 'S' & 'T' denote the *source* and *target* transformation functions respectively, and the subscripts 'sn' & 'tn' denote the Gaussian approximation of the magnitude spectra of the two noises being compared. $F_{T_{avg}}$ and $F_{S_{avg}}$ are the average formant frequencies of the *target* and *source* languages respectively. The modification of mean is contributed to by both language and noise transfer. The modification in standard deviation is controlled by noise transfer alone.

5 Experiments and Results

The pipeline of the experiments conducted started with the generation of a Trapezoidal Delta function through CLPSO for a certain combination of Language, Noise type and SNR level. This led to a certain improvement in Short Time Objective Intelligibility (STOI) (Taal et al., 2010) for that specific combination. Thereafter transfer was done across languages and the results were compared with that obtained through direct training through CLPSO. Next, the CLPSO was used to obtain the Gaussian Delta function for a combination as discussed before. Thereafter, transfer was done across noises, followed by a combination of language and noise. These results can be exemplified by one of the cases of training and transfer, as shown in Table 1. The table compares the CLPSO training of both Trapezoidal as well as Gaussian Delta functions, followed by the transfer across Languages and Noises respectively. The final column of the Table shows the results of the universal TL framework working on Languages and Noises simultaneously. It can be seen that the CLPSO training performs better using Gaussian than Trapezoid. Also transfer across Noises improves intelligibility more than transfer across Languages for this particular setup. Furthermore, the universal TL results seem to approach the results as obtained through direct training.

6 Future Plans and Road-map for Thesis

The work done up to this point has provided certain insightful results. These results will be used as prior information to train a Neural Network to develop a fully autonomous system that is scalable in terms of applications. Thus, the thesis is projected to be consisting of two parts. The former dealing with intelligibility improvement of speech using basic machine learning and optimization, while the latter handling these issues through the employment of deep neural networks.

References

Karan Nathwani, Morgane Daniel, Gaël Richard, Bertrand David, and Vincent Roussarie. 2016. Formant shifting for speech intelligibility improvement in car noise environment. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5375–5379. IEEE.

Meysam Rahmati, Reza Effatnejad, and Amin Safari. 2014. Comprehensive learning particle swarm optimization (clpso) for multi-objective optimal power flow. *Indian Journal of Science and Technology*, 7(3):262–270.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217. IEEE.