



# Automatic Estimation of Intelligibility Measure for Consonants in Speech

Ali Abavisani<sup>1</sup>, Mark Hasegawa-Johnson<sup>1</sup>

<sup>1</sup>Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign

aliabavi@illinois.edu, jhasegaw@illinois.edu

## Abstract

In this article, we provide a model to estimate a real-valued measure of the intelligibility of individual speech segments. We trained regression models based on Convolutional Neural Networks (CNN) for stop consonants /p,t,k,b,d,g/ associated with vowel /a/, to estimate the corresponding Signal to Noise Ratio (SNR) at which the Consonant-Vowel (CV) sound becomes intelligible for Normal Hearing (NH) ears. The intelligibility measure for each sound is called SNR<sub>90</sub>, and is defined to be the SNR level at which human participants are able to recognize the consonant at least 90% correctly, on average, as determined in prior experiments with NH subjects. Performance of the CNN is compared to a baseline prediction based on automatic speech recognition (ASR), specifically, a constant offset subtracted from the SNR at which the ASR becomes capable of correctly labeling the consonant. Compared to baseline, our models were able to accurately estimate the SNR<sub>90</sub> intelligibility measure with less than 2 [dB<sup>2</sup>] Mean Squared Error (MSE) on average, while the baseline ASR-defined measure computes SNR<sub>90</sub> with a variance of 5.2 to 26.6 [dB<sup>2</sup>], depending on the consonant.

**Index Terms:** speech perception in noise, human speech recognition, objective intelligibility measures.

## 1. Introduction

The primary purpose of hearing aids is to improve speech perception. But the speech signal has little role in tuning current hearing aid technologies. There has been no consensus on how to involve speech in the procedure. In clinical audiology, it is usual that hearing impaired (HI) patients complain about their difficulty in understanding speech in noisy environments. A prerequisite for proper advice regarding the patient's ability to communicate in noisy situations, or for selection of the optimal hearing aid amplification, is a reliable clinical test to assess patient's speech perception in noise. However, developing such a test is a very complicated task due to the large number of factors that are involved in the measurements [1]. Thus, hearing speech in background noise should be a substantial part of a clinical audiology test to assess hearing loss (HL). If the spectrum of the masker sound is shaped according to the long-term average of the speech signal, the test results will be less dependent on the speaker [1].

Psycho-acoustic speech recognition experiments with human subjects using Consonant-Vowel (CV) sounds as speech stimulus have a long history [2], and can therefore be effectively calibrated. Since about 58% of the phonetic segments in spoken English are consonants [3], consonant recognition scores are appropriate for the evaluation of speech intelligibility. Recorded CV stimuli vary in their intelligibility, however: some stimuli that are clearly intelligible under quiet conditions become unin-

telligible with only a small amount of added noise, apparently because of the presence of conflicting cues for the place of articulation [4]. In order to be useful as a test of HL audibility thresholds, it is necessary to select CV speech stimuli that are intelligible to NH listeners at the test SNR.

In normal hearing ears each consonant becomes masked at a token dependent threshold, denoted SNR<sub>90</sub>. The SNR<sub>90</sub> is defined as the SNR at which NH ears can recognize the token correctly, with at least 90% probability, averaged across NH listeners. As the noise is increased from Quiet (no noise), the identification of most sounds goes from less than 0.5% error to 10% error (at SNR<sub>90</sub>), and then to chance performance, over an SNR range of just a few [dB] (i.e., less than 10 [dB]) [5]. Hence SNR<sub>90</sub> is an important token-specific threshold metric of noise robustness. Since SNR<sub>90</sub> is based on NH perception, it is a perceptual measure of understanding speech in noise, for a NH ear. It would be very useful if one could assign an SNR<sub>90</sub> label to each speech token so in a speech-based test with background noise, the audiologist would know which tokens are appropriate for speech perception assessment. Additionally, as the SNR<sub>90</sub> is a measure corresponding to the intensity of the primary cue region [6], knowledge of this perceptual measure could be used to enhance speech playback in noisy environments: after measuring the background noise level, the playback device could amplify each syllable as necessary to guarantee that every segment is played at an SNR higher than its own SNR<sub>90</sub>.

Previous studies show that, at SNRs well above SNR<sub>90</sub>, HI listeners will have errors in recognizing a small subset of CV stimuli out of all the presented stimuli [7, 8]. Once high error sounds have been identified, one may seek the optimum treatment (insertion gain) for a patient's hearing aid.

In this study, we propose a model to estimate the SNR<sub>90</sub> for CV speech sounds based on CNN (we use only 1D convolution, with an architecture based most closely on [9]). The model is a supervised estimator of SNR<sub>90</sub> in dB. Particularly, current work is focused on SNR<sub>90</sub> estimation for stop consonants /p, t, k, b, d, g/ in association with vowel /a/. To accomplish this goal, one needs a suitable dataset of CV speech sounds to train the model. One obstacle is that examining each CV with 30 NH listeners in psycho-acoustic speech recognition experiments takes a tremendous amount of time, and only a handful of sounds can be evaluated this way. A CNN trained using such a small corpus does not achieve low error rates. To overcome this challenge, we propose a speech augmentation method by manipulating speech characteristics in ways that do not affect the consonant recognition score for an average NH listener. These manipulations include inducing microphone characteristics (high pass or low pass effect with small attenuation), pitch shift (up and down) and consonant duration manipulation (compression and extension).

This study is a novel approach to incorporate speech in the

process of tuning hearing aids, using machine learning. Previous works that use machine learning to assist the design of hearing aids are mainly focused on using deep learning to estimate amplification gain [10], or suppress noise and reverberation for speech enhancement [11]. The models proposed here may be used in the audiology clinic to propose perceptual stimuli for hearing aid fine tuning.

The task proposed in this paper is to estimate the SNR at which any particular spoken syllable becomes intelligible to an NH listener. To the best of our knowledge, there is no published algorithm that performs the same task, therefore there is no published baseline to which the results of this study can be compared. In order to create a baseline for comparison, therefore, we use the SNR at which a commercial ASR (Google cloud’s speech to text model [12]) becomes capable of correctly transcribing the same consonant. Since the threshold SNR for ASR success is always much higher than the threshold SNR for human listeners, our baseline measure is computed by subtracting a constant offset from the ASR SNR.

Section 2 explains the psycho-acoustic experiments used to determine the  $SNR_{90}$  of each token. In section 3, we describe the procedures for speech augmentation and how to generate appropriate labels for distorted sounds. The CNN-based model to estimate the intelligibility measure is described in section 4. Results and discussion will follow in section 5.

## 2. $SNR_{90}$ determination

To determine the  $SNR_{90}$  of target tokens, we presented them to 30 NH listeners, aged 19-30 years old, at various SNRs ranging from -22 to 22 [dB]. Hearing thresholds were tested prior to the experiments to rule out any hearing impairments. Listeners were given 14 buttons, labeled with 14 consonants of English, and were asked to select the consonant they heard. The speech signal was mixed with speech-weighted noise as described by [13] to set the SNR to -22, -18, -12, -6, 0, 6, 12, 18 and 22 [dB] respectively. Presentation order was randomized over consonant, talker, and SNR.

The experiment was designed using a two-down-one-up strategy: if the subject recognizes the token correctly, the SNR drops two levels [12dB], otherwise it increases one level [6dB]. This schedule is consistent with conventional paradigms in audiology testing. If the subject loops between two consecutive SNRs at least three times, the presentation concludes for that token.

After collecting the data from all NH listeners, we average the response accuracies for each token at each SNR. For the  $n^{\text{th}}$  subject at each SNR, the probability of correct response for the  $c^{\text{th}}$  token is calculated as:

$$P_c(n, SNR) = \frac{N_{\text{correct}}(n, SNR)}{N_{\text{total}}(n, SNR)} \quad (1)$$

where  $N_{\text{correct}}(n, SNR)$  is the number of correct recognitions at the specified SNR, and  $N_{\text{total}}(n, SNR)$  is the total number of tokens presented at that SNR, for the  $n^{\text{th}}$  subject. Hence, the average score is:

$$\overline{P_c}(SNR) = \frac{1}{N_{\text{sub}}} \sum_{n=1}^{N_{\text{sub}}} P_c(n, SNR) \quad (2)$$

where  $N_{\text{sub}}$  is the number of subjects.

The plot of average accuracy ( $\overline{P_c}(SNR)$ ) versus SNR was (in our data) always greater than 90% at SNRs above  $SNR_{90}$ , and always less than 90% at SNRs below  $SNR_{90}$ . In order to estimate the exact value of  $SNR_{90}$  for each token, we linearly interpolated between the smallest  $\overline{P_c}(SNR)$  above 90% and the value of  $\overline{P_c}(SNR)$  at the next lower SNR, and then measured the SNR at which the linear interpolation crosses 90%; Fig. 1 shows an example of this procedure. As can be observed in Fig. 1, All NH listeners recognized the /ba/ sound whose results are schematized in Fig. 1 with no error above 0 [dB] SNR. At  $SNR=-6$  [dB], subjects started to have some errors, but still correctly recognized the CV with accuracy better than 90%. When the SNR further drops to -12 [dB],  $\overline{P_c}(SNR)$  suddenly drops below 50%. Linear interpolation of  $\overline{P_c}(SNR)$  estimates that for this /ba/ sound the  $SNR_{90}$  is -6.66 [dB].

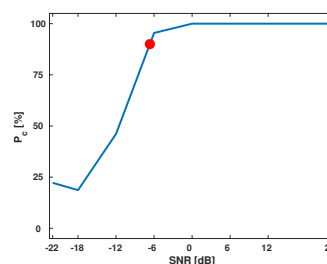


Figure 1: An example of determining the  $SNR_{90}$  of a /ba/ token; the  $SNR_{90} = -6.66$  [dB] is shown with a red dot and is determined by linear interpolation along the  $P_c$  averaged across 30 NH listeners.

Using this procedure, we examined 14 consonants associated with vowel /a/, spoken by 16 different talkers, both male and female (total of  $14 \times 16 = 224$  tokens), to determine their  $SNR_{90}$ . To train our model, we selected tokens with  $SNR_{90}$  below -3 [dB] to focus on tokens with better intelligibility measure. To have balanced examples from various talkers, we limited our training data to female talkers and stop consonants. Hence, from the pool of 224 evaluated tokens, we selected 39 tokens to build the model ( $8 \times /ka/$ ,  $7 \times /ta/$ , 6 of each other stop consonants). Table 1 provides the contributing tokens along with their  $SNR_{90}$  from experiments with NH listeners.

Table 1: CV tokens that are used as original undistorted tokens for training the model to estimate intelligibility measure. For each CV, its  $SNR_{90}$  in [dB] is provided.

Talker	CV	$SNR_{90}$	CV	$SNR_{90}$	CV	$SNR_{90}$
f101	/ka/	-5	/ba/	-11		
	/pa/	-17	/ta/	-21	/ka/	-11
f103	/ba/	-3	/da/	-17	/ga/	-13
	/pa/	-13	/ta/	-17	/ka/	-8
f105	/ba/	-9	/da/	-21	/ga/	-11
	/pa/	-11	/ta/	-11	/ka/	-12
f106	/ga/	-5				
	/pa/	-11	/ta/	-21	/ka/	-4
f108	/ba/	-11	/da/	-12	/ga/	-3
	/pa/	-4	/ta/	-17	/ka/	-4
f109	/ba/	-11	/da/	-12	/ga/	-4
	/ta/	-17	/ka/	-9	/ba/	-4
f113	/da/	-18				
	/pa/	-11	/ta/	-18	/ka/	-5
f119	/da/	-18	/ga/	-6		

### 3. Speech augmentation

The number of presentations of each CV, during each perceptual experiment, depends on the number of trials necessary to find the SNR at which the subject makes mistakes (Sec. 2), but on average, each token requires a total listening time of three minutes. Each SNR<sub>90</sub> measurement is the result of 30 NH listeners, a total of 90 minutes/token. To train a viable CNN-based SNR<sub>90</sub> estimation model for each CV, one would need thousands of different versions of such CV sound, each with a measured SNR<sub>90</sub>. Instead of manually labeling such a large number of training tokens, we began with only 39 labeled tokens, and introduced various distortions using methods that have previously been shown to have little effect on the SNR<sub>90</sub> of human listeners. In this way, it is possible to generate sufficient data to be able to train the model.

#### 3.1. Applied distortions

The distortions applied in the current study include extending and compressing the duration of the consonant, shifting the pitch of the whole CV up and down, and introducing channel effects by applying low pass and high pass filtering with small attenuation. We used Praat [14] for pitch and duration manipulations, and MATLAB<sup>®</sup> for filtering. Each token (i.e., a CV sound with a specific talker) was distorted according to every available single distortion, but not any combinations. Table 2 provides specifics of distortions applied to each CV sound.

Table 2: Various distortions applied to each tested CV sound for speech augmentation.

Distortion		Details
Duration change	Extend	from 1:1 up to 1:3 in steps of 0.3%
	Compress	from 1:1 down to 1:0.5 in steps of 1%
Pitch Shift	Up	up to 600 [Hz] in steps of 1 [Hz]
	Down	down to 20 [Hz] in steps of 1 [Hz]
FIR Filter (200 degree)	High pass	20 cut-offs log-spaced between 0.2-3 [kHz], 20 attenuations ranged between 0.6-12 [dB]
	Low pass	20 cut-offs log-spaced between 1-8 [kHz], 20 attenuations ranged between 0.6-12 [dB]

The artificially distorted CV token might have a different SNR<sub>90</sub> than the original token. The possible changes in SNR<sub>90</sub> are controlled by conducting new psycho-acoustic experiments with NH subjects, similar to the experiment described in section 2, to measure the SNR<sub>90</sub> of the most distorted token along each distortion continuum. If the measured SNR<sub>90</sub> is greater than 6 [dB], we did not include the entire sequence of tokens generated by such distortion in our augmented speech dataset. Table 3 provides the tokens that are not valid after applying various distortions.

For each CV sound, we may assume that every token on the continuum, between the original unmodified token and the most distorted token, will have an SNR<sub>90</sub> that is somewhere between the original token’s SNR<sub>90</sub> and the most distorted token’s SNR<sub>90</sub>. Thus, for each distortion scheme, we linearly interpolated between the SNR<sub>90</sub> of the original and most distorted token in order to generate SNR<sub>90</sub> labels for tokens in between. Although this procedure may not produce the exact experimentally accurate SNR<sub>90</sub> measure for every token, it generates approximate SNR<sub>90</sub> labels that are useful as training data. Using these methods, the original 39 unmodified tokens were expanded to

Table 3: CV tokens (vowel /a/ is omitted) from various talkers that have SNR<sub>90</sub> greater than 6 [dB] for their most distorted case, thus were not included in the training data.

Talker	Duration		Pitch Shift		Filtering	
	Extend	Compress	Up	Down	LPF	HPF
f101		/k/		/k/	/k/	/k/
f103	/b/	/k,b,g/	/b/		/b,g/	
f105	/d/	/p,k/	/p,d,g/		/k/	/p,k,g/
f106	/t/	/t,k/			/t/	
f108	/b,d,g/	/k,g/	/k,b,g/		/k,g/	/k/
f109	/p,g/	/p,t,k,d,g/	/k,d,g/		/t,k,g/	
f113	/k,d/	/k/	/b/		/k/	
f119	/p,k,d,g/	/k,g/	/k,g/	/k/	/k,g/	/k,g/

create 43201 tokens (/pa/: 8121 tokens, /ta/: 9726 tokens, /ka/: 6267 tokens, /ba/: 7065 tokens, /da/: 7977 tokens, and /ga/: 4045 tokens). Different talkers contributed different token counts for each consonant. This is sufficient CV speech data to train the model for SNR<sub>90</sub> estimation.

### 4. Model structure

The dataset of unmodified CV tokens contain recorded *wav* files naturally spoken by various talkers (table 1). After speech augmentation, the dataset increased to 43201 *wav* files of CV tokens. Each file has a duration of less than three seconds and contains an isolated CV utterance with a sampling frequency of 16 [kHz]. The data was divided to train, development and test partitions with non-overlapping talkers, to train, tune and test the individual model for each stop consonant. The percentage of partitions of data was different for various stop consonants, as each talker contributed differently in the final dataset after augmentation. The exact number of tokens for training, development and test sets are provided in table 4.

Table 4: Total number of tokens ( $N_{Total}$ ) after speech augmentation, along with the number of allocated tokens for training ( $N_{Train}$ ), development ( $N_{Dev}$ ) and test ( $N_{Test}$ ), to train models for SNR<sub>90</sub> estimation.

CV	$N_{Total}$	$N_{Train}$	$N_{Dev}$	$N_{Test}$
/pa/	8121	6261	1298	562
/ta/	9726	7787	974	965
/ka/	6267	4966	703	598
/ba/	7065	5622	981	462
/da/	7977	5910	1129	938
/ga/	4045	3004	569	472

For each CV, the time interval from the start of the consonant till the end of the onset of vowel /a/ is manually segmented. Within this interval, the 320 point log magnitude Short-Time Fourier Transform (STFT) with 75% overlapping Hamming windows of length 25 [msec] is extracted to feed into the input layer.

For each of the sounds /pa/, /ta/, /ka/, /ba/, /da/ and /ga/, we trained a separate model to estimate the SNR<sub>90</sub>. The models are based on Convolutional Neural Networks (CNN) [9], which include convolutional layers that act in the time domain. The model contains 3 to 7 convolutional layers with Rectified Linear Unit (ReLU) nonlinear activation function [15], reduced to the input of the fully connected layer (FC) by average pooling, followed by two fully connected layers with ReLU non-linearity in the hidden layer and a linear output node that

produces the estimated  $\text{SNR}_{90}$  value. The loss function is Mean Squared Error (MSE) between estimated  $\text{SNR}_{90}$  and correct  $\text{SNR}_{90}$  label. We used stochastic gradient descent to minimize the loss. The model is implemented in TensorFlow 1.4 [16]. To avoid over fitting, dropout [17] is applied to the fully connected layer, with dropout rates tuned on the development set. Table 5 illustrates the common structure of the network for various stop consonants. The differences between the models for different CVs are in their hyper-parameters.

Table 5: Common structure of the network trained for various CVs. The number of convolutional layers, and the time domain kernel values of  $[w_1, w_2, w_3]$  in convolutional layers, are among the hyper-parameters trained for each CV model separately, and are reported in table 6; conv4-7 are extra layers added during tuning.

Layer	Kernel	(stride, pad)	Input	Output
conv1	$1 \times w_1 \times 320 \times 128$	(1,0)	log  STFT	conv1
conv2	$1 \times w_2 \times 128 \times 256$	(1,0)	conv1	conv2
conv3	$1 \times w_3 \times 256 \times 512$	(1,0)	conv2	conv3
conv4-7	$1 \times w_3 \times 512 \times 512$	(1,0)	conv3-6	conv4-7
FC	$512 \times 1024$	-	conv3-7	FC
out	$1024 \times 1$	-	FC	$\text{SNR}_{90}$

The hyper-parameters for each model are trained by using the development data. These hyper-parameters include number of convolutional layers, time domain kernel size for each layer, learning rate for gradient descent optimization, batch size, and dropout rate. Table 6 provides the parameters for each model after fine-tuning with development data. In table 6,  $N_{\text{CNN}}$  indicates the number of convolutional layers,  $[w_1, w_2, w_3]$  refers to the time domain kernel size in convolutional layers,  $N_{\text{Batch}}$  indicates the batch size, and  $\eta$  indicates the learning rate. If the network has more than three convolutional layers, the time domain kernel size beyond the third layer is set equal to  $w_3$ .

Table 6: Hyper-parameters tuned separately for each stop consonant model.

CV	$N_{\text{CNN}}$	$[w_1, w_2, w_3]$	$N_{\text{Batch}}$	$\eta$	$P_{\text{dropout}}$
/pa/	3	[5, 7, 7]	8	$10^{-4}$	50%
/ta/	7	[7, 3, 3]	4	$10^{-4}$	17%
/ka/	3	[7, 5, 7]	4	$10^{-5}$	38%
/ba/	3	[3, 3, 3]	16	$10^{-4}$	10%
/da/	7	[5, 5, 3]	8	$10^{-6}$	33%
/ga/	3	[5, 3, 7]	4	$10^{-6}$	8%

## 5. Results

Human speech perception data from experiments with NH listeners form the ground truth for the evaluation of automatic estimates of  $\text{SNR}_{90}$ . To compare the human perception of CV tokens versus machine perception, we tested several commercial ASRs, and chose the one with the best performance for these data, Google cloud’s speech to text interface [12]. The phone call model in Google’s speech to text system is an enhanced model that aims to have better performance in noisy environments.

Since the speech to text system is trained to recognize words and sentences, it is unable to recognize non-word CVs, therefore we counted, as correct, any word containing the target consonant followed by a non-high vowel, averaged over 40 trials for each CV. Output not containing the target CV, and empty output transcript, were both counted as failure to recognize.

Table 7 provides the comparison between human perception of CV sounds versus ASR for stop consonants in the test corpus. The lowest SNR at which the output transcript of the ASR contained a word including the target CV is reported as the ASR estimate of  $\text{SNR}_{90}$ . The average perceptual evaluation of speech quality (PESQ) score [18] is also measured and reported for CV sounds at the ASR  $\text{SNR}_{90}$ . In table 7, the  $\text{SNR}_{90}$ s from human and ASR, and the PESQ scores are averaged across different talkers (indicated by  $N_T$ ) for the same CV.

Table 7:  $\text{SNR}_{90}$  (in [dB]) of CV stimuli, average results across  $N_T$  talkers, measured by human subjects (ground truth) and ASR. The PESQ score is calculated at ASR  $\text{SNR}_{90}$ . Bias (in [dB]) and variance ( $\sigma^2$ ) of ASR  $\text{SNR}_{90}$  estimation (in  $[\text{dB}^2]$ ), as well as our model’s test MSE (in  $[\text{dB}^2]$ ) are provided.

CV	$N_T$	$\text{SNR}_{90}$		PESQ	ASR		Our model Test MSE
		Human	ASR		Bias	$\sigma^2$	
/pa/	6	-11.2	2.4	2.35	13.6	26.6	1.71
/ta/	7	-17.4	0.4	2.17	17.8	16.5	1.45
/ka/	8	-6.7	4.2	2.47	10.9	12.4	1.29
/ba/	6	-7.4	4.2	2.54	11.6	16.5	1.71
/da/	6	-16.3	-4.5	2.09	11.8	14.2	1.81
/ga/	6	-7	4.5	2.4	11.5	5.2	1.89

Variance of the ASR estimated  $\text{SNR}_{90}$  ranged from 5.2  $[\text{dB}^2]$  (/ga/) to 26.6  $[\text{dB}^2]$  (/pa/), with an average of 15.3  $[\text{dB}^2]$ . In comparison, our CNN-based models were able to estimate the  $\text{SNR}_{90}$  of various CV sounds with small errors. The mean squared error of estimation for test sounds were all below 2  $[\text{dB}^2]$  for the models trained for each consonant. Table 7 illustrates the test errors in  $\text{SNR}_{90}$  estimation for stop consonants.

## 6. Conclusion

In this study, we introduced new models based on convolutional neural networks to estimate the  $\text{SNR}_{90}$  of individual speech stimuli.  $\text{SNR}_{90}$  is defined to be the SNR at which normal hearing listeners are able to correctly recognize a stimulus with 90% probability, and has been shown to be related to the level of the primary cue to consonant identity [13]. One important application of such models is to evaluate various speech sounds before using them to assess speech perception in humans, e.g., to tune hearing aids. The main advantage of using the models developed here is to estimate intelligibility of speech syllables in background speech-weighted noise, without the need of running expensive and time consuming experiments with human subjects in controlled conditions. The speech augmentation methods introduced in the current study help to increase the size of the training database adequately to train deep learning models for speech processing. Our results show that the developed models outperform the only available baseline, namely, the SNR at which an ASR is able to correctly recognize each consonant. Variance of the ASR-based estimate of  $\text{SNR}_{90}$  is 15.3  $[\text{dB}^2]$  (with a bias greater than 11 [dB]), while MSE of the deep-learning-based estimator is below 2  $[\text{dB}^2]$ .

## 7. Acknowledgments

This work utilizes resources supported by the National Science Foundation’s Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign.

## 8. References

- [1] A. W. Bronkhorst and R. Plomp, "A clinical test for the assessment of binaural speech perception in noise." *Audiology*, 29(5), 275-285, 1990.
- [2] G. A. Miller and P. E. Nicely, "Analysis of Perceptual Confusions Among Some English Consonants." *J. Acoust. Soc. Am.* 27, 338-352, 1955.
- [3] M. A. Mines, B. F. Hanson and J. E. Shoup, "Frequency of occurrence of phonemes in conversational English." *Language and speech*, 22(3), 221-241, 1978.
- [4] F. Li, A. Menon, and J. B. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech." *J. Acoust. Soc. Am.*, 127(4):2599-2610, 2010.
- [5] J. C. Toscano and J. B. Allen, "Across-and within-consonant errors for isolated syllables in noise." *Journal of Speech, Language, and Hearing Research*, 57(6), 2293-2307, 2014.
- [6] A. Kapoor and J. B. Allen, "Perceptual effects of plosive feature modification." *J. Acoust. Soc. Am.*, 131(1):478-491, 2012.
- [7] A. Abavisani and J. B. Allen, "Evaluating hearing aid amplification using idiosyncratic consonant errors." *J. Acoust. Soc. Am.*, 142(6):3736-3745, 2017.
- [8] A. C. Trevino and J. B. Allen, "Within-consonant perceptual differences in the hearing impaired ear." *J. Acoust. Soc. Am.*, 134(1):607-617, 2013.
- [9] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang, "Phoneme recognition using time-delay neural networks." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3), 328-339, 1989.
- [10] S. I. M. M. R. Mondol and S. Lee, "A Machine Learning Approach to Fitting Prescription for Hearing Aids." *Electronics*, 8(7), 736, 2019.
- [11] Y. Zhao, D. Wang, I. Merks and T. Zhang, "DNN-based enhancement of noisy and reverberant speech." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6525-6529, 2016.
- [12] Google Inc., "Google Cloud Speech-to-text." <https://cloud.google.com/speech-to-text>, 2020.
- [13] S. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise." *J. Acoust. Soc. Am.*, 121(4):2312-26, 2007.
- [14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]." <http://www.praat.org>, Version 6.1.12, 2020.
- [15] X. Glorot, A. Bordes and Y. Bengio, "Deep sparse rectifier neural networks." *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315-323, 2011.
- [16] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et. al., "Tensorflow: A system for large-scale machine learning." *12th Symposium on Operating Systems Design and Implementation*, pp. 265-283, 2016.
- [17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research*, 15(1), pp. 1929-1958, 2014.
- [18] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs." *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Cat. No. 01CH37221), Vol. 2, pp. 749-752, 2001.