# Perceptual Effects of Nasal Cue Modification

Fan Bai[1,2,*]

[1]*Department of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, P.R. China*

[2]*Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, USA*

**Abstract:** Acoustic or perceptual cues used for speech perception can be very helpful in almost all areas of speech signal processing. A new methodology 3-Dimensional-Deep Search and a new visualized intelligible time-frequency computer-based model AI-gram have been developed and are being researched since the last several years (Human Speech Recognition (HSR) research group at the University of Illinois Urbana-Champaign) for isolation of stable perceptual cues of consonants. The perceptual cues of nasal consonants [1] have been successfully found considering these techniques [1]. The previous work is extended by assessing the changes in nasal sound perception and cue region is modified by using digital signal processing method. The amplitude of the perceptual cue region is amplified, attenuated or ignored completely and then the perception score is measured. A high correlation between the amplitude of the cue region and the perception score is found. The intelligibility of the entire token is increased or decreased approximately in a similar fashion as the cue region modified amplitude which is measured by the MMSE shift of the perceptual score curve. This validates that the regions identified are perceptual cue regions for nasal consonants. The digital signal processing method proposed can be used as a new approach for enhancing speech signal in noisy conditions.

**Keywords:** Acoustic cue, modification, nasal cue, nonlinear spectrum analyzer, speech perception.

## 1. INTRODUCTION

The cochlea is a nonlinear spectrum analyzer. Once a speech sound reaches the cochlea, it is represented by time varying energy patterns across the basilar membrane (BM). However, only a small subset of the patterns called perceptual cues contribute to speech recognition. Identification of perceptual cues can be very helpful in almost all areas of speech signal processing. A psychoacoustic method named the 3-Dimensional-Deep Search (3DDS) [2] has been developed for obtaining real human speech acoustic cues of consonants in the past five years by the Human Speech Recognition (HSR) research group at the University of Illinois Urbana-Champaign.. It combined three independent psychoacoustic experiments with human natural speech as stimuli and paired with a new computer-based time-frequency model, the AI-gram [3-5], which simulates human auditory peripheral processing and predicts the audibility of speech with introduction of masking noise by showing the audible parts on a time-frequency graph. Its name AI-gram has been derived from the well-known speech Articulation Index (AI), developed by Fletcher [6]. The block diagram for AI-gram is shown in Fig. (**1**).

The objective of 3DDS method is to measure the significance of speech subcomponents on perception in three dimensions: time, frequency, and intensity as shown in Fig. (**2**). The various parts of a speech sound along the three axes are systematically removed; truncated in time; high-/low-pass filtered in frequency; or masked with white noise. The effect of the removed component due to the change in the respective recognition score is assessed. Finally, the acoustic (or perceptual) cue of stop consonants [2], fricative consonants [7] and nasal consonants [1] are successfully located by 3DDS technique.

The acoustic cue of nasal consonant /m/ lying in the onset of the F2 formant region in the previous research ranges between 0.35 and 1.2 kHz as highlighted by the red rectangle in the AI-gram in Fig. (**3A**) and the acoustic cue of /n/ lying in an F2 transition region around 1.5 kHz is also highlighted by the red rectangle in the AI-gram in Fig. (**3B**). The research presented here is an attempt to extend the previous work to explore the effect of changing cue strength on the perception of the sound. For plosive consonants, similar research has been carried out by Allen and Li (2009) [8]; Li and Allen (2011) [9]; Li *et al.* (2010) [2]; Régnier and Allen (2008) [3].

## 2. METHODS

### A. Speech Stimuli

Natural speech tokens are obtained from the University of Pennsylvania Linguistic Data Consortium “Articulation Index Corpus” (LDC2005S22). More details for this corpus can be found in a study by Li *et al*. [2]. These tokens include

*Address correspondence to this author at Department of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, P.R. China; Tel: +001 2174189782;
E-mail: baifan11111@gmail.com

1874-1290/15 2015 Bentham Open

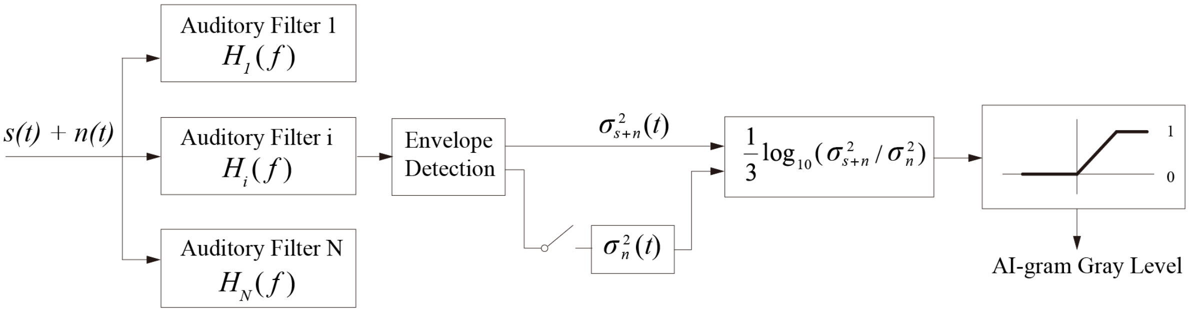

s(t) + n(t)
Auditory Filter 1 $H_1(f)$
Auditory Filter i $H_i(f)$
Auditory Filter N $H_N(f)$
Envelope Detection
$\sigma^2_{s+n}(t)$
$\sigma^2_n(t)$
$\frac{1}{3}\log_{10}(\sigma^2_{s+n}/\sigma^2_n)$
1
0
AI-gram Gray Level

**Fig. (1).** Block diagram for how an AI-gram is computed (modified according to Li *et al.*, 2010).

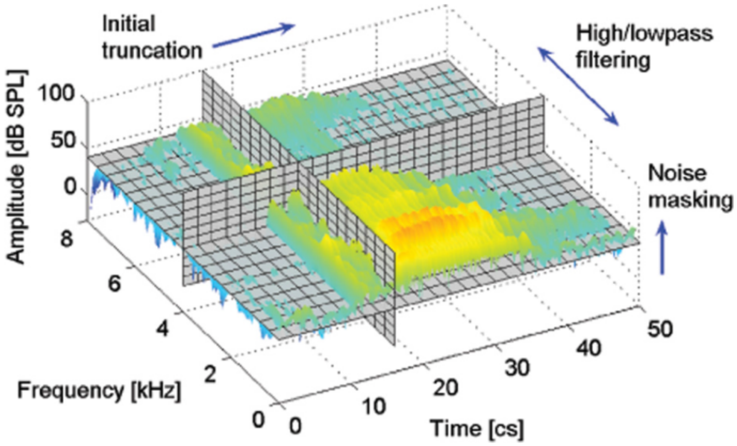
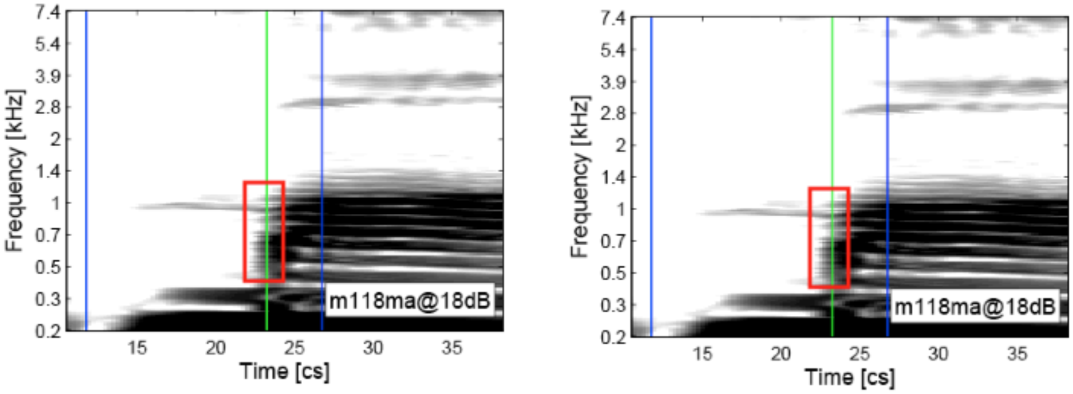


**Fig. (2).** The 3DDS method for acoustic cue identification [7].



**(A)** **(B)**

**Fig. (3).** Perceptual cue of the nasal consonants /m/ (**A**) and /n/ (**B**).

/ma/ and /na/ CV syllables spoken by eight different talkers, half male and half female; and /pa, ba, fa, da, ka, sa, va, ga, za/ CV syllables spoken by six different talkers, half male and half female. For controlling variability, the /ma, na/ tokens are exactly the same as used in previous experiments. The other nine CV syllables are added in order to prevent the subjects from deducing the experimental subset. All speech stimuli are sampled at 16 kHz with 16 bit analog to digital converter.

### B. Modification to Stimuli

The cue region of each token is first identified by 3DDS technique as listed in Table **1**. Each cue region is defined by the starting time point ($t_{start}$), duration ($\Delta t$), lower frequency boundary ($F_{lo}$) and upper frequency boundary ($F_{hi}$). Then these regions are modified by amplification, attenuation and complete removal.

**Table 1. Information on each sound.**

| Sound | $t_{start}$ | $\Delta t$ | $F_{lo}$ | $F_{hi}$ | $\Delta SNR_{+}$ | $\Delta SNR_{-}$ |
|---|---|---|---|---|---|---|
| | [cs] | [cs] | [Hz] | [Hz] | [dB] | [dB] |
| f103ma | 17 | 2.5 | 363 | 1300 | -4.7 | 4.4 |
| m102ma | 31 | 2 | 363 | 1400 | -3.8 | 6.1 |
| m115ma | 30 | 2.5 | 363 | 1250 | -8.1 | 7.3 |
| f105ma | 29.5 | 1 | 363 | 1250 | -7.1 | 6.4 |
| m118ma | 23 | 1 | 363 | 1250 | -7.7 | 7.3 |
| f113ma | 24.5 | 2 | 363 | 1250 | -5.5 | 5.9 |
| f119ma | 19.5 | 1.5 | 363 | 1150 | -4.3 | 6.1 |
| m120ma | 28.5 | 2 | 363 | 1250 | -7.5 | 6.9 |
| f113na | 32 | 4 | 1350 | 2164 | -5.5 | 6.1 |
| f109na | 28.5 | 3 | 1250 | 2164 | -6.4 | 5.4 |
| m112na | 29 | 7 | 1100 | 1649 | -3.5 | 3.9 |
| m102na | 28.5 | 3 | 1100 | 1649 | -5.9 | 5.3 |
| f101na | 23.5 | 4.5 | 939 | 1649 | -5.8 | 3.6 |
| m120na | 25 | 4.5 | 1100 | 2164 | -4.9 | 6.3 |
| f105na | 26.5 | 4 | 1250 | 1649 | -5.7 | 6.2 |
| m118na | 29 | 4.5 | 1250 | 2164 | -5.0 | 3.8 |

All the cue regions listed in Table **1** are modified. To achieve this, the original speech signal is transformed from time domain into frequency domain by using short-time Fourier transform (STFT) as represented in equation (1):

$$X\left[m,k\right]=\sum_{n=0}^{N-1}w\left[n\right]s\left[mR-n\right]e^{-j2\pi kn/N} \tag{1}$$

Where *s[n]* represents the original speech, *w[n]* is a Kaiser window with a −91 dB sidebands attenuation, which means that first side lobe is 91 dB smaller than the main lobe. The length of each overlapping frame is 20-ms with 5-ms containing R new samples. This can be deduced from equation (1) that the original speech signal is time-reversed and then shifted to Kaiser window with a step of 5-ms new samples.

The coefficients of STFT *X[m, k]* in time-frequency domain are calculated prior and then the cue regions are? (AUTHOR: Some information is missing here)

After multiplying this gain matrix, the modified speech spectrum is obtained as follows:

$$Y\left[m,k\right]=X\left[m,k\right]\cdot M\left[m,k\right] \tag{2}$$

The modified signal in time-frequency domain is converted back to the time domain by using an inverse Fourier transform:

$$y\left[m,n\right]=\frac{1}{N}\sum_{k=0}^{N-1}Y\left[m,k\right]e^{j2\pi kn/N} \tag{3}$$

The Overlap Add (OLA) synthesis [10] is applied at the end and the resultant modified speech signal *y[n]* is represented as:

$$y\left[n\right]=\sum_{m=-M_0}^{0}y\left[mR,n\right] \tag{4}$$

### C. Noise Conditions

After modification, white noise is added to both the unmodified and modified speech stimuli at seven different SNRs: -18, -15, -12,-6, 0, 6, and 12 dB. The SNRs are based on the unmodified speech stimuli, so the same noise added to original speech stimuli is added to all the other modified speech stimuli. Examples of the AI-gram after modification and noise addition are shown in Figs. (**4** and **5**).

### D. Procedure

Stimuli are diotically presented to subjects *via* Sennheisser HD 280 Pro headphones ranging between 75 and 80 dB sound pressure level. Sound pressure level is calibrated at 1 kHz tone using Radioshack sound pressure level meter. The experiments are conducted in a sound-proof booth and silence is also ensured outside the booth to proceed experiments without causing any interruption. The total number of tokens is 826 (2 nasal consonants × 8 talkers × 7 SNRs × 4 conditions + 9 other consonants × 6 talkers × 7 SNRs = 826 tokens). Each token is only presented once in a random fashion. A GUI interface using MATLAB is developed to display 13 options for the subjects to choose from. These options include two nasal consonants /m, n/, the nine other consonants/p, b, f, d, k, s, v, g, z/, and two additional options, “Noise Only” and “Others”. After listening to the sound stimuli, the subjects respond by clicking one of the 13 buttons labeled with one of these 13 options. There is also a repeat button for the subjects to listen each token up to 3 times before responding.

A mandatory practice session is held before each experimental session. In the practice session, after subjects respond to the stimuli, the correct sound appears on the GUI interface. If the subject answers correctly, the sound would be removed from the play list; otherwise it would be put back to the list randomly, until they can correctly recognize it. In the experimental session, no feedback is considered for avoiding learning effects.

### E. Subjects

The study had been carried out on 20 participants primarily undergraduate students with normal sense of hearing. All of them were under 40 and self-reported, having no hearing disorder. A pilot test showed that under 12dB SNR, all participants correctly recognized the two nasals /m/ and /n/. Remuneration has also been provided for participation.

## RESULTS AND ANALYSIS

The experiment data is first recorded in the form of confusion patterns, which means the proportion of all responses for a particular token is a function of SNR [11]. Then the

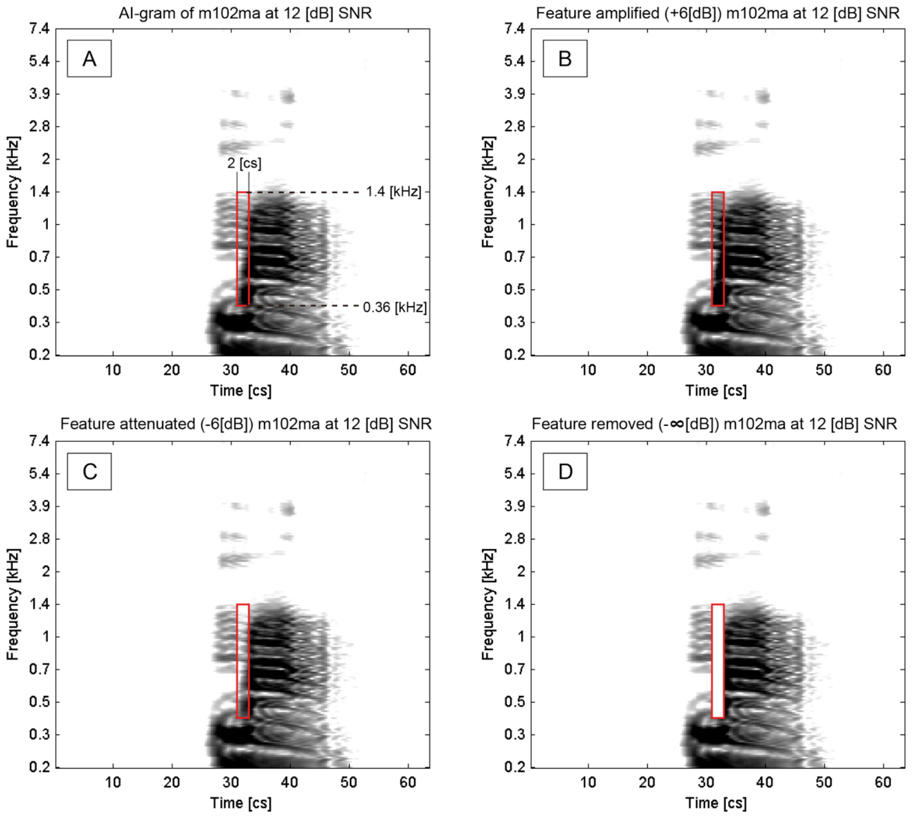


**Fig. (4).** AI-gram of the unmodified (**A**), cue-amplified (**B**), cue-attenuated (**C**), and cue-removed token m102ma.

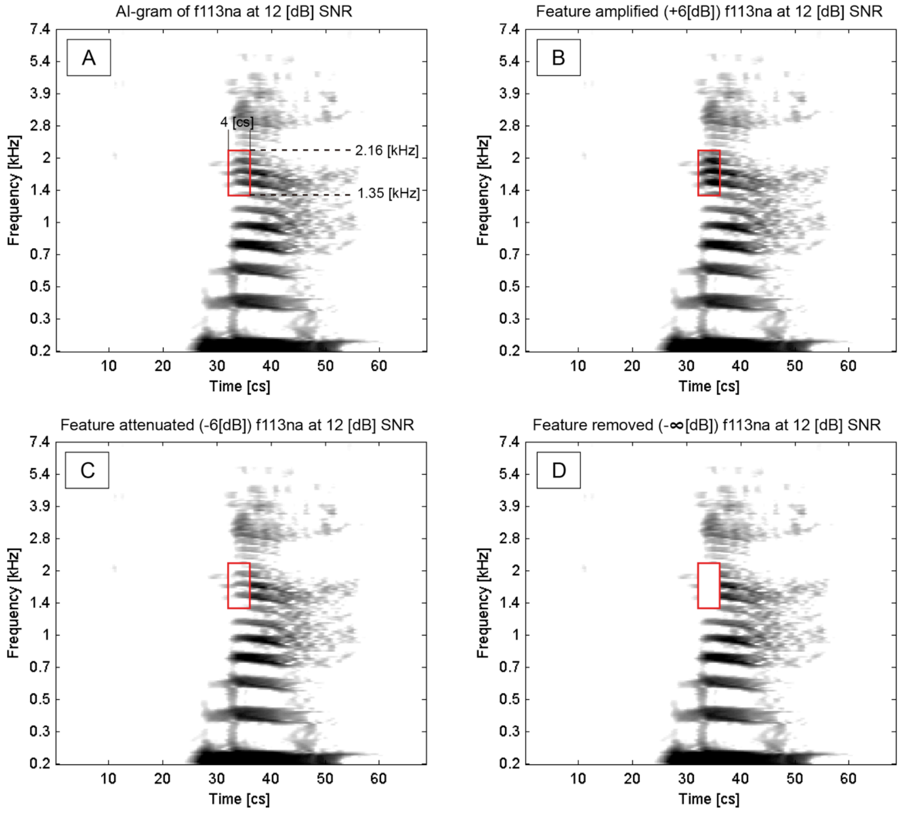


**Fig. (5).** AI-gram of the unmodified (**A**), cue-amplified (**B**), cue-attenuated (**C**), and cue-removed token f113na.modified by multiplying the corresponding coefficients with a two-dimensional gain matrix. Specifically, the gain of cue removal is - ∞ dB corresponding to 0, the gain of cue attenuation is -6 dB corresponding to 1/2, and the gain of cue amplification/enhancement is + 6 dB corresponding to 2.

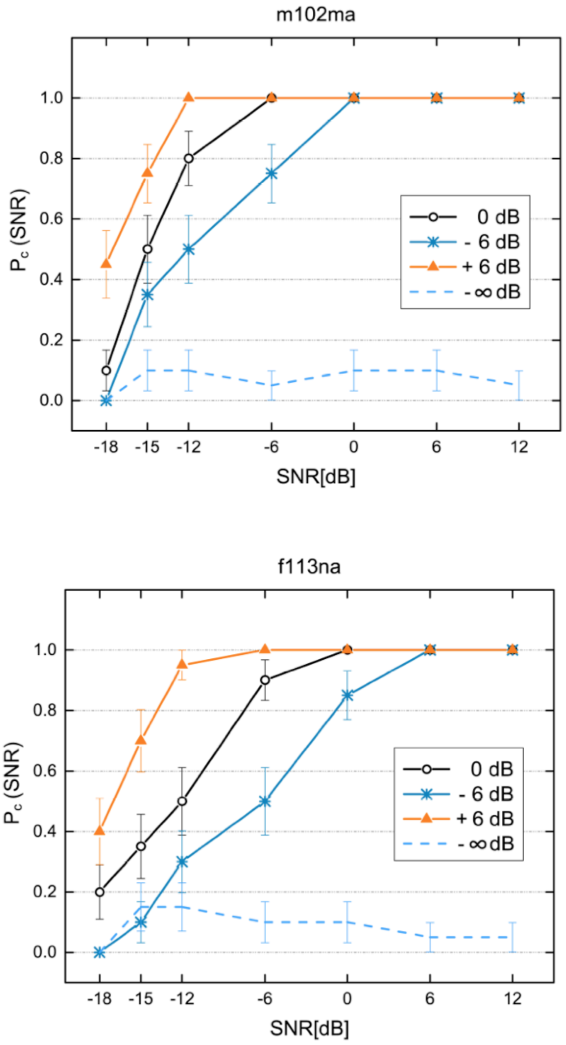


**Fig. (6).** Comparison of recognition scores of the original and the three modifications (cue-amplified, cue attenuated and cue-removed) of sounds m102ma and f113na.

recognition score, or perceptual score (i.e. the correct perceptual probabilities) for each token $P_c$ and the error rate $P_e$ are calculated. Again, both $P_c$ and $P_e$ vary as a function of SNR. For easier observation and comparison of recognition scores across the modified and original sounds for each token, these are shown as a line plot. In Fig. (**6**), the line plots for token m102ma and f113na are plotted. It is evident from the plot that the curves of the modified tokens are shifted compared to the corresponding curve of the original unmodified token. In fact, the effectiveness of the modification is indicated by the amount of shift. The cue-amplified token always shifts to the left along the axis of SNR as compared to the perceptual score curve of the original token. This means that the sound becomes more robust and can be recognized at lower SNR for getting the same perceptual score as the corresponding original token. Conversely, the perceptual score curve of cue-attenuated token shifts to the right, indicating that for getting the same perceptual score, the listeners need to hear it under higher SNR than the original token. The curve of cue-removed token has lowest performance as expected. In this case, almost all listeners cannot correctly recognize cue-removed token at any of the SNRs. The standard deviation bars at each data point are calculated assuming that experiment procedure behaves as Bernoulli trials, where the token is recognized either correctly or incorrectly by different listeners. So the standard deviation, where N is the number of responses for each token at each SNR and is also equal to the number of listeners.

For quantitative evaluation of the effectiveness of the modification, for each of the cue-amplified, cue-attenuated and original token, their recognition scores are assigned to a sigmoid function. It has been shown that the sigmoid reasonably estimates the average $P_c$ for CVs [12]. More specifically,as illustrated in Fig. (**7**), for each token, first the error rate is assigned under each SNR, denoted as $P_e(SNR)$, to the sigmoid function:

$$P_e\left(SNR\right) = \frac{e_c}{1+e^{\lambda\left(SNR-SNR_0\right)}} \qquad (5)$$

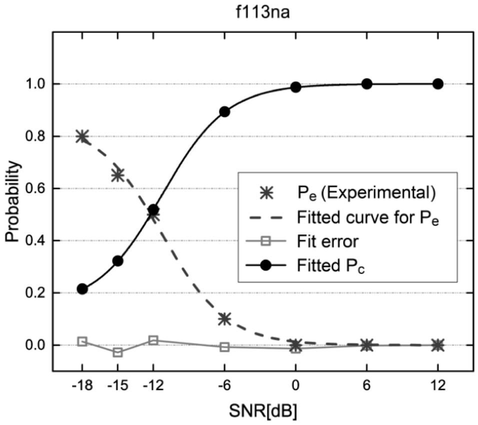


**Fig. (7).** Regression analysis of recognition score for the unmodified token f113na.

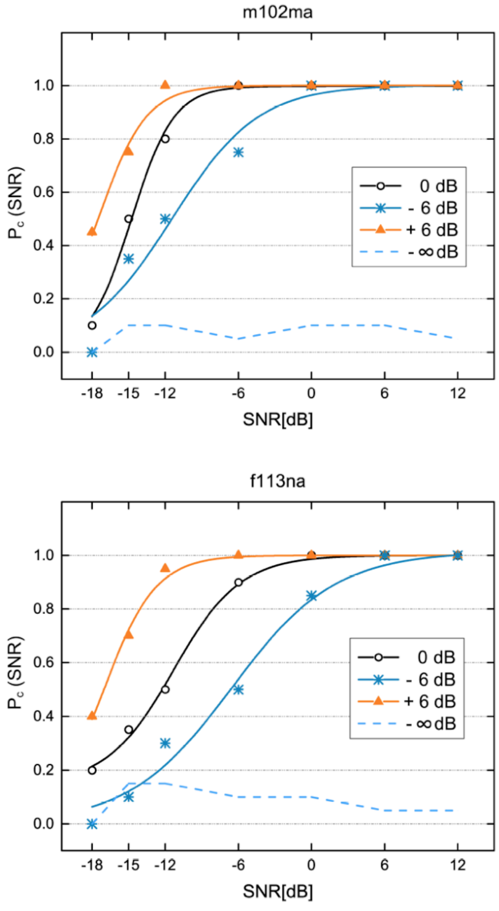


**Fig. (8).** Comparison of sigmoid fits for modified and original sound, using token m102ma and f113na as examples.

Here $e_c$ is the probability of error at chance, which is the error rate by guessing randomly from the candidate CV choices. In this case, $e_c = 10/11$ as there are 11 CV choices. $\lambda$ and $SNR_0$ are the two parameters to be determined. $\lambda$ is the scaling factor ranging between 1 and 0. $SNR_0$ is the speech recognition threshold at which 50% of the speech can be correctly recognized. Then the curve is generated for the correct perceptual probability under each SNR: $P_c(SNR) = 1\text{-}P_e(SNR)$. The sigmoid fits for the two example tokens with the original and three modified versions are shown in Fig. (**8**).

To estimate the overall lateral shift for the modified token, the minimum mean square error (MMSE) calculation is used. Each modified sigmoid is shifted with step size of 0.01, until the mean squared difference between the shifted and the unmodified sigmoid is minimized (Fig. **9**). The ΔSNR estimated for each token is listed in Table **1**.

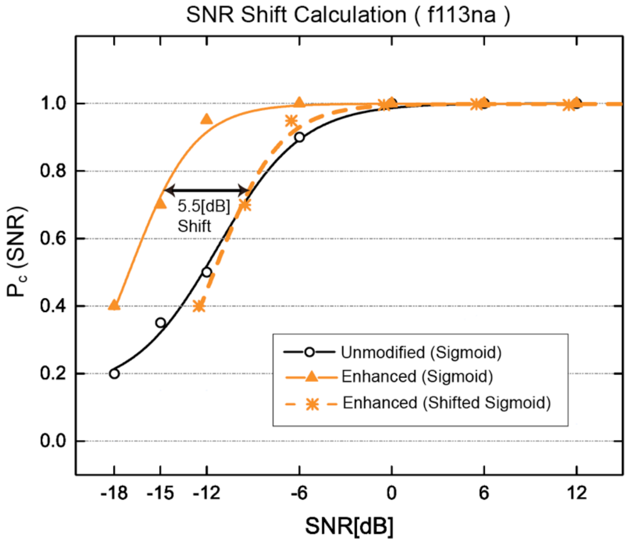


**Fig. (9).** Calculation of the sigmoid shift for token f113na.

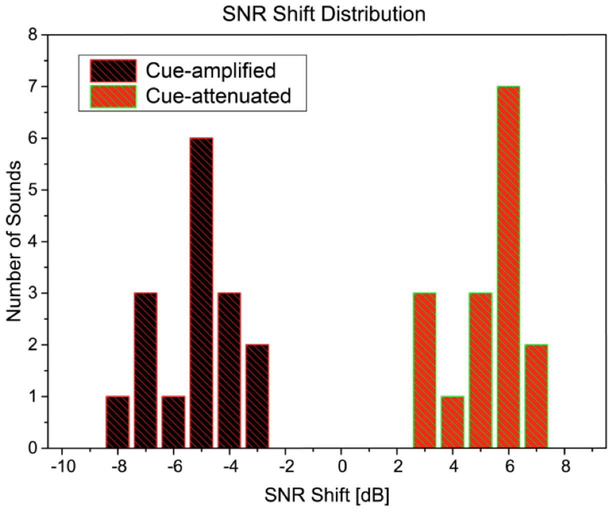


**Fig. (10).** Histogram of the SNR shift for all cue-amplified and cue-attenuated curves.

The results indicate that these cue regions are in fact important regions for perception of nasal consonants. There is obvious shift in the perceptual score curve for both the types of modifications in each of the tokens as shown in Figs. (**6** and **8**). From -6 dB to -18 dB, there is also almost no overlap in the standard deviation bars indicating the three curves for the cue-amplified, the original and the cue attenuated sound having almost no overlap in their distribution. This result demonstrates that the modification in these small regions alone can significantly change the perceptual score.

Second, these cue regions could be carrying the majority of perceptual information. As shown in Figs. (**10** and **11**), 29 out of the total 32 tokens (2 nasal × 2 modification types (amplified and attenuated) × 8 talkers) showed more than 4 dB lateral shift, while only 3 tokens (f103ma, f112na, f101na) demonstrated absolute shifts less than 4 dB. The average SNR shift of amplified version for /m/ and /n/ in dB is -6.09 and -5.34, with standard deviation of 1.71 and 0.89 respectively; the average SNR shift of attenuated version for /m/ and /n/ in dB is 6.30 and 5.08, with standard deviation of 0.94 and 1.14 respectively (Table **2**). These data demonstrate that modifying a small cue region shifts the entire original curve by around 6 dB. This number is very close to the actual amount of cue amplification/ attenuation. Thus, the magnitude of the MMSE shift of the original curve in dB (i.e. the magnitude of the relative SNR change compared to the original sound) is very similar to the magnitude of the cue amplitude modification. In other words, an increase of 6 dB in the cue region alone has the same effect as an increase of 6 dB in SNR for the whole token, while a decrease of 6

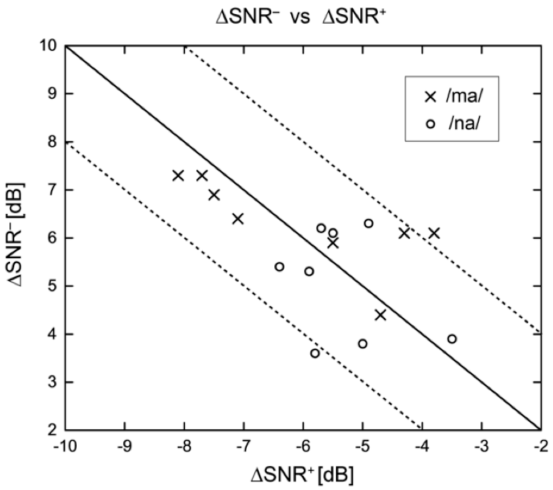


**Fig. (11).** SNR shift for cue-amplification versus cue-attenuation of each token.

dB in the cue region alone has the same effect as a decrease of 6 dB in SNR for the whole token. Together with the fact that completely removes the cue region results in almost 0 perceptual score even at 18 dB SNR (dashed line, Figs. (**6** and **8**), it is indicated that the region modified contains key perceptual component – the perceptual cue.

**Table 2. Mean and standard deviation of perceptual curve shift for cue-amplified and cue-attenuated nasal sounds.**

| | /ma/ | | /na/ | |
|---|---|---|---|---|
| | ΔSNR+[dB] | ΔSNR– [dB] | ΔSNR+[dB] | ΔSNR– [dB] |
| Mean | -6.09 | 6.30 | -5.34 | 5.08 |
| Standard Deviation | 1.71 | 0.94 | 0.89 | 1.14 |

Thus, these results validate speech cues identified previously for the two nasals.

## CONCLUSION

In this study, firstly, the amplitude modification of the perceptual cue region of the nasal consonants located by the 3DDS method has been carried out. Secondly, the perception score of three versions of cue modification (amplification, attenuation, and removal) has been measured. The results depicted high correlation between the amplitude of the modified cue region and the recognition score. More specifically, when the region was amplified, the recognition score increased correspondingly; when the amplitude of the region attenuated, the recognition score decreased accordingly; when the region was removed, the token could not be recognized even under very high SNR. Furthermore, results show that the intelligibility of the entire token, as represented by the perceptual score curve as a function of SNR, is increased and decreased approximately in a similar magnitude as the cue amplitude modification, based on the MMSE shift measures. This result provided new supporting evidence for the conclusion of previous research on nasal perceptual cue, that the perceptual cue of /ma/ is the onset of the F2 formant region between 0.35 and 1.2 kHz ; and that the perceptual cue of /na/ is an F2 transition region around 1.5 kHz. It can also be used as the proof that normal hearing listeners recognize consonants depending on these perceptual cues under a wide range of SNRs. The digital signal processing method proposed here can be used as a new way of enhancing speech signal for noisy environments.

## CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

## ACKNOWLEDGEMENTS

The author would like to acknowledge the help provided by Jont Allen, Feipeng Li, Andrea Trevino, and other members of the Human Speech Research group at the University of Illinois. This research was partially supported by a University of Illinois grant.

## REFERENCES

[1] F. Li, *Perceptual Cues of Consonant Sounds and Impact of Sensorineural Hearing Loss on Speech Perception*, University of Illinois at Urbana-Champaign, Urbana, IL, 2009.

[2] F. Li, A. Menon, and J.B. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2599-2610, 2010.

[3] M.S. Régnier, and J.B. Allen, "A method to identify noise-robust perceptual features: Application for consonant /t/", *The Journal of the Acoustical Society of America*, vol. 123, pp. 2801-2814, 2008.

[4] B.E. Lobdell, *Models of Human Phone Transcription in Noise Based on Intelligibility Predictors*, University of Illinois at Urbana-Champaign, Urbana, IL, 2009.

[5] B.E. Lobdell, J.B. Allen, and M.A. "Hasegawa-Johnson, Intelligibility predictors and neural representation of speech", *Speech Communication*, vol. 53, no. 2, pp. 185-194, 2011.

[6] J.B. Allen, "How do humans process and recognize speech?", *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 567-577, 1994.

[7] F. Li, A. Trevino, A. Menon, and J. B. Allen, "A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise", *The Journal of the Acoustical Society of America,* vol. 132, no. 4, pp. 2663–2675, 2012.

[8] J.B. Allen, and F. Li, "Speech perception and cochlear signal processing", *IEEE Signal Processing Magazine*, vol. 26, pp. 73-77, 2009.

[9] F. Li, and J. B. Allen, "Manipulation of Consonants in Natural Speech", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 496-504, 2011.

[10] J.B. Allen, "Short time spectral analysis, synthesis, and modification by discrete Fourier transform", *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-25, no.3, pp. 235-238, 1977.

[11] J.B. Allen, "Consonant recognition and the articulation index", *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2212-2223, 2005.

[12] N.R. French, and J.C. Steinberg, "Factors governing the intelligibility of speech sounds", *The Journal of the Acoustical Society of America,* vol. 19, pp. 90-119, 1947.

Received: February 20, 2015 Revised: March 26, 2015 Accepted: April 03, 2015