Modeling the level-dependent changes of concurrent vowel scores^{a)}

Harshavardhan Settibhaktini^{b)} and Ananthakrishna Chintanpalli

Department of Electrical and Electronics Engineering, Birla Institute of Technology and Science,
 Pilani Campus, Vidya Vihar, Pilani, Rajasthan, 333031, India

(Received 12 July 2017; revised 20 October 2017; accepted 2 January 2018; published online 26 January 2018)

The difference in fundamental frequency (F_0) between talkers is an important cue for speaker segregation. To understand how this cue varies across sound level, Chintanpalli, Ahlstrom, and Dubno [(2014). *J. Assoc. Res. Otolaryngol.* **15**, 823–837] collected level-dependent changes in concurrent-vowel identification scores for same- and different- F_0 conditions in younger adults with normal hearing. Modeling suggested that level-dependent changes in phase locking of auditory-nerve (AN) fibers to formants and F_0 s may contribute to concurrent-vowel identification scores; however, identification scores were not predicted to test this suggestion directly. The current study predicts these identification scores using the temporal responses of a computational AN model and a modified version of Meddis and Hewitt's [(1992). *J. Acoust. Soc. Am.* **91**, 233–245] F_0 -based segregation algorithm. The model successfully captured the level-dependent changes in identification scores of both vowels with and without F_0 difference, as well as identification scores for one vowel correct. The model's F_0 -based vowel segregation was controlled using the actual F_0 -benefit across levels such that the predicted F_0 -benefit matched qualitatively with the actual F_0 -benefit as a function of level. The quantitative predictions from this F_0 -based segregation algorithm demonstrate that temporal responses of AN fibers to vowel formants and F_0 s can account for variations in identification scores across sound level and F_0 -difference conditions in a concurrent-vowel task.

© 2018 Acoustical Society of America. <https://doi.org/10.1121/1.5021330>

[ICB]

Pages: 440–449

I. INTRODUCTION

Younger adults with normal hearing (YNH) have a remarkable ability to segregate two or more simultaneous speech sounds when presented at varying sound levels. They utilize many cues to identify the speech during this multi-talker scenario. Some of these cues are the differences in onset and offset timing of the speech sounds, differences in speech characteristics (e.g., formant differences), and differences in fundamental frequency (F_0) between the speakers (e.g., Cherry, 1953; Brox and Nootboom, 1982; Zwicker, 1984; Bregman, 1990). Among these, F_0 difference is widely studied as an important cue for segregating multiple speech signals, as F_0 is physiologically unique to a given individual. Concurrent vowel identification is often studied to understand how F_0 difference helps to identify two simultaneously presented vowels. In this experimental paradigm, two vowels with equal duration and level are presented simultaneously to one ear of a human listener. The task of the listener is to identify both vowels, and performance is measured as a function of F_0 difference between the two vowels. For YNH listeners, identification scores for both vowels improve as F_0 difference increases between the two vowels, and then asymptotes usually at ~ 3 -Hz F_0 difference

or higher (Assmann and Summerfield, 1990; Summerfield and Assmann, 1991; Culling and Darwin, 1993; Arehart *et al.*, 1997, 2005; Summers and Leek, 1998; Vongpaisal and Pichora-Fuller, 2007; Chintanpalli and Heinz, 2013; Chintanpalli *et al.*, 2016). However, these studies have been typically conducted at a single vowel level.

To understand how the ability to utilize the F_0 difference cue varies across levels, Chintanpalli *et al.* (2014) collected concurrent-vowel data for 0- and 26-Hz F_0 difference conditions (i.e., same and different F_0 , respectively) in YNH listeners (generally between 20 and 26 years). Figure 1 shows the percent correct identification scores as a function of vowel level for same and different F_0 conditions (modified from Chintanpalli *et al.*, 2014). Their subjects showed an improvement in percent identification score as vowel level was increased from low-to-medium [25–50 dB sound pressure level (SPL)] and then a decline was observed at higher levels (65–85 dB SPL). The F_0 benefit is defined as the difference in percent correct identification scores between the different and same F_0 conditions, and is commonly used in the concurrent-vowel literature (e.g., Assmann and Summerfield, 1990; Summers and Leek, 1998; Arehart *et al.*, 2005; Chintanpalli *et al.*, 2014, 2016). The mean percent of F_0 benefit increased from 25 to 50 dB SPL and then remained fairly constant from 50 to 85 dB SPL.

Chintanpalli *et al.* (2014) further performed computational modeling by quantifying the phase locking of auditory-nerve (AN) fibers to vowel formants and F_0 s using the average localized synchronized rate (Young and Sachs, 1979) and template

^{a)}A portion of this research work was presented at the 40th meeting of Association for Research in Otolaryngology, Baltimore, MD, USA, February 2017.

^{b)}Electronic mail: s.harsha@pilani.bits-pilani.ac.in

FIG. 1. Percent identification scores of both the vowels for same F_0 (squares) and different F_0 (triangles) as a function of the vowel level. Note that the figure shown is the percent identification scores rather than the rationalized arcsine transformed scores, modified from Chintanpalli *et al.* (2014). Error bar denotes ± 1 standard error of the mean. The F_0 benefit is indicated by an arrow.

contrast (Larsen *et al.*, 2008), respectively. These two neural coding schemes were computed from the responses of a well-regarded AN model (Zilany *et al.*, 2009). It was inferred that the F_0 -difference cue was degraded at 25 dB SPL because this level had the lowest F_0 benefit, which could be attributed to poor phase locking (or rate place coding) of AN fibers to F_0 s. Furthermore, the reduced scores at higher levels were attributed to poor phase locking to formants (particularly to the second formant), resulting from broader auditory filters at higher SPLs. It was assumed that the phase locking of AN fibers to formants and F_0 s contribute separately to vowel identification. The listeners first try to segregate the vowel pair (based on the available cues), and then identify them individually. This suggests that the vowel segregation precedes identification, and thus one needs to incorporate the interactions between formant and F_0 -difference cues for identification. However, the separate analyses of phase locking to formants and F_0 s limits the direct conclusions that can be made from the modeling work of Chintanpalli *et al.* (2014) in terms of the level-dependent changes in identification scores of both vowels. Additionally, their conclusions were based on two concurrent-vowel pairs (/a, æ/ and /i, æ/) and did not compute any of the actual identification scores. Hence, this limited modeling approach only allowed inferences to be made that the level-dependent changes in neural coding schemes (i.e., average localized synchronized rate and template contrast) may have sufficient information to predict the concurrent-vowel data. These conclusions would have been stronger had this study computed the actual identification scores.

The current study significantly extends the modeling work used in Chintanpalli *et al.* (2014) to test explicitly their conclusions, which were based on neural coding schemes. Here, we compute the actual identification scores using an improved version of the AN model (Zilany *et al.*, 2014) and a modified version of a well-established F_0 -based segregation algorithm (Meddis and Hewitt, 1992). This segregation

algorithm takes into account the interactions between formant and F_0 difference cues, and computes the identification scores for concurrent vowels. A similar type of modeling framework has also been used in previous studies to successfully capture (at least qualitatively) the effect of F_0 difference on vowel identification for a given level (Meddis and Hewitt, 1992; Chintanpalli and Heinz, 2013). The aim of the present study was to examine whether this modeling framework can successfully predict the level-dependent changes in identification scores of both vowels across various sound levels and for same- and different- F_0 conditions.

II. METHODS

A. Stimuli

The stimulus generation for the current study was similar to those reported in Chintanpalli *et al.* (2014). A set of five different vowels (/i/, /u/, /a/, /æ/ and /ɜ:/) were generated using a MATLAB implementation of a cascade formant synthesizer (Klatt, 1980). The duration of each vowel was 400 ms, including 15-ms raised-cosine rise and fall ramps. The formant frequencies and the bandwidth of each vowel are shown in Table I, which is similar to the earlier studies on concurrent vowel identification (e.g., Assmann and Summerfield, 1994; Summers and Leek, 1998; Chintanpalli and Heinz, 2013; Chintanpalli *et al.*, 2014, 2016).

A concurrent vowel was obtained by adding any two individual vowels. To form a vowel pair for the different F_0 condition, one vowel had a constant $F_0 = 100$ Hz, whereas another vowel had $F_0 = 126$ Hz. These five vowels were arranged in different combinations to obtain 25 concurrent vowels. To maintain equivalent numbers of vowel pairs when F_0 difference was 0 Hz, this condition also included 25 vowel pairs (five identical-vowel pairs + ten different-vowel pairs presented twice, where F_0 was 100 Hz for both vowels in the pair). A total of 50 concurrent vowels (25 vowel pairs $\times$ two F_0 conditions) was used at each level. The individual vowel ranged from 25 to 85 dB SPL and the vowel pairs either had the same (100 Hz) or different F_0 s (100 and 126 Hz). Overall, 300 vowel pairs were generated (25 vowel pairs $\times$ 2 F_0 conditions $\times$ 6 levels).

B. Computational modeling: Predicting identification scores across sound level and F_0 -difference conditions

To understand the level-dependent changes underlying the peripheral processing associated with the identification scores of both vowels (Fig. 1), the current study utilizes the

TABLE I. Formants in Hz for five different vowels. Values in parenthesis of first column correspond to bandwidth around each formant (in Hz).

Vowel	/i/	/a/	/u/	/æ/	/ɜ:/
F1 (90)	250	750	250	750	450
F2 (110)	2250	1050	850	1450	1150
F3 (170)	3050	2950	2250	2450	1250
F4 (250)	3350	3350	3350	3350	3350
F5 (300)	3850	3850	3850	3850	3850

computational model (see Fig. 2) by cascading an AN model (Zilany *et al.*, 2014) with a modified version of F_0 -based segregation algorithm (Meddis and Hewitt, 1992). The Meddis and Hewitt (1992) F_0 based segregation algorithm is the only algorithm that has successfully captured (at least qualitatively) the effect of F_0 difference on concurrent vowel identification.

1. Auditory-nerve (AN) model

The AN model developed by Zilany *et al.* (2014) was used to predict the AN responses to concurrent vowels. It is an extension of several previous versions of the model, which have been tested extensively against neurophysiological responses from cats to pure tones, two-tone complexes, broadband noise, and vowels (Carney, 1993; Zhang *et al.*, 2001; Heinz *et al.*, 2001; Bruce *et al.*, 2003; Tan and Carney, 2003; Zilany and Bruce, 2006, 2007, Zilany *et al.*, 2009). Relevant to the current study, this model captures (1) the level-dependent changes in cochlear nonlinearities, e.g., compression, suppression, broadened tuning, and best-frequency shifts with increase in sound levels (e.g., Ruggero *et al.*, 1997), and (2) the level-dependent changes in phase locking ability of AN fibers to vowel formants (e.g., Miller *et al.*, 1997; Zilany and Bruce, 2006, 2007; Chintanpalli *et al.*, 2014) and to F_0 coding (e.g., Chintanpalli *et al.*, 2014).

The input to the AN model was a concurrent vowel and the output was a time varying discharge rate of a single AN fiber for a particular characteristic frequency (CF). The AN responses were computed from 100 different logarithmically spaced CFs, ranging from 125 Hz to 4000 Hz, for each of the

three fiber types (see step 1 of Fig. 2). As per the anatomical and physiological functions of AN fiber, there are three fiber types that constitute the AN population. These fiber types are high spontaneous rate (HSR), medium spontaneous rate (MSR), and low spontaneous rate (LSR) (Liberman, 1978). The overall discharge rate at each CF was the weighted sum of the discharge rates as per the distribution of the spontaneous rates (HSR = 0.61, MSR = 0.23, and LSR = 0.16; Liberman, 1978; see step 1 of Fig. 2), which is relevant for modeling the level effects for concurrent vowels since AN-fiber threshold is inversely related to its spontaneous rate (SR).

2. F_0 -based segregation algorithm for concurrent-vowel identification

To understand the changes in identification score of both vowels across F_0 difference, there have been many attempts to develop physiologically inspired computational models that could validate the effect of F_0 difference on the identification scores of both the vowels (Assmann and Summerfield, 1990; Meddis and Hewitt, 1992; Chintanpalli and Heinz, 2013). The Meddis and Hewitt (1992) model was the first to successfully capture the effect of F_0 difference on concurrent-vowel identification. Subsequently, Chintanpalli and Heinz (2013) replicated this effect using the same F_0 -based segregation algorithm of Meddis and Hewitt (1992) but with a more recent AN model (Zilany and Bruce, 2007). However, the parameter values of this segregation algorithm were different in each study because of the different peripheral models that were used.

In this segregation algorithm (see step 2 of Fig. 2), an auto-correlation function (ACF) was computed for each of

FIG. 2. Block diagram illustrating the steps involved in computational modeling for predicting the level-dependent changes in concurrent vowel scores for same- and different- F_0 conditions. The first stage is the responses from the AN model, whereas the second stage is the modified version of the F_0 -based segregation algorithm for vowel identification.

the weighted sum of the time-varying discharge rate of AN fibers. Each ACF was multiplied by a single exponential function with a time constant ($\tau = 10$ ms, as in Meddis and Hewitt, 1992). A pooled-auto-correlation function (PACF) was computed by summing across the ACFs from different fibers. The dominant F_0 was estimated by identifying the peak value in the pitch region (4.5–12.5 ms) of the PACF. The algorithm then identifies the number of individual ACFs that had a peak at the dominant F_0 . If the identified channels were greater than the user-defined segregation parameter, then the algorithm decided there was one F_0 ; otherwise two F_0 s were present. For the latter condition, the algorithm proceeded to vowel segregation, i.e., the entire pool of ACFs were divided into two different sets (e.g., ACF1: CFs that showed a peak at the dominant F_0 and ACF2: CFs that did not show a peak at the dominant F_0). The segregated PACF1 and PACF2 were computed separately. The individual vowels were identified based on the inverse Euclidean distance of the timbre region (0.1–4.5 ms) of the segregated PACF with an individually stored timbre regions of single vowels. The single vowel template was obtained by averaging the timbre regions of the PACFs across 6 F_0 conditions (i.e., 100, 101.3, 103, 106, 112, and 126 Hz). If a single F_0 was predicted, then the vowel segregation was achieved only using the formant difference cues. The inverse Euclidean distances were computed between the timbre region of the combined PACF and the timbre regions of the individually stored templates of single vowels. The algorithm predicted a single vowel (identical vowels) only when the ratio of the best-to-second best-match was greater than the user defined parameter (i.e., $m1/m2$), otherwise the two different vowels were predicted.

A modified version of the Meddis and Hewitt (1992) segregation algorithm was used in the current study with the following changes. First, CF-dependent time constant was used for computing the ACFs across CFs (Cariani, 2004; Bernstein and Oxenham, 2005; Chintanpalli *et al.*, 2014). This might account for the effect of peripheral filtering on pitch perception. More specifically, smaller-bandwidth peripheral filters (i.e., at lower CFs) result in longer-duration impulse responses and thus require slower time constants. Higher-bandwidth peripheral filters (i.e., at higher CFs) result in shorter-duration impulse responses and thus require faster time constants. The value of τ was varied per CF ($\tau = 50$ ms for $100 \leq \text{CF} < 440$ Hz; $\tau = 36$ ms for $440 \leq \text{CF} < 880$ Hz; $\tau = 30$ ms for $880 \leq \text{CF} < 1320$ Hz; $\tau = 29$ ms for $\text{CF} \geq 1320$ Hz). Second, a single-vowel template at each level (“optimal case”) was obtained by averaging the timbre regions of the PACFs for 100 and 126 Hz. The rationale for using a separate template for each sound level was mainly derived from the level-dependent changes due to cochlear nonlinearities.

III. RESULTS

In concurrent-vowel data (see Fig. 1), the F_0 benefit increased with increasing vowel level (25–50 dB SPL) and then remained fairly constant from 50 dB SPL onwards. This suggests that the ability to segregate vowels using F_0

difference cue also increases up to 50 dB SPL. Thus, this criterion was used in the current modeling to constrain the user-defined segregation parameter of the Meddis and Hewitt (1992) algorithm across sound level. More specifically, this parameter was selected such that (1) percent segregation ability in the two- F_0 condition, computed across 25 vowel pairs, remained almost constant from 50 dB SPL onwards, and (2) percent segregation is zero at each level for same- F_0 condition. Table II shows the absolute values of the segregation parameter and its corresponding percentage F_0 segregation at each level. To further constraint the model, $m1/m2$ parameter was fixed to 2.5 across vowel levels.

Figure 3(A) shows the model scores for both vowels as a function of level for same (squares) and different F_0 (triangles) conditions. The identification score improved as vowel level increased from low-to mid-levels and declined at higher levels for both F_0 conditions. These patterns of identification scores are qualitatively similar to that of listeners’ identification scores (see Fig. 1).

Figure 3(B) shows the model’s percent vowel segregation, computed across pairs, as a function of vowel level for same (squares) and different- F_0 (triangles) conditions. As expected, the percent segregation was zero for same- F_0 condition across all levels [squares in Fig. 3(B)]. This indicates that the level-dependent changes in identification scores for same F_0 could be influenced largely by changes in cochlear nonlinearities (e.g., bandwidth of auditory filter changes with level), which may then affect the level-dependent changes in the availability of formant-difference cues for identification. The percent segregation increased for different F_0 [triangles in Fig. 3(B)] until 50 dB SPL, in order to match the actual F_0 benefit of the concurrent-vowel data, at least qualitatively.

Figure 3(C) shows the model scores for percent correct identification of one vowel of the pair as a function of vowel level (solid line). The identification score was 100% and was independent of F_0 -difference and vowel-level conditions. For comparison, this figure also shows the actual percent correct identification score of one vowel as a function of vowel level, calculated from the concurrent-vowel data (dashed line). The model score was successful in capturing the identification score of one vowel, quantitatively. These findings suggest that F_0 difference and vowel level are primarily important for identifying the second vowel of the pair [compare Figs. 3(A) vs 3(C), solid line or Figs. 1 vs 3(C), dashed line]. Additionally, it may suggest that the level-dependent changes in identification scores of both vowels

TABLE II. The absolute values of the segregation parameter and its corresponding percent F_0 segregation at each vowel level.

Level (dB SPL)	User defined segregation parameter	Percentage of F_0 segregation
25	36	48
35	43	56
50	60	84
65	52	76
75	52	76
85	56	80

FIG. 3. Effect of level on concurrent-vowel identification for same- and different- F_0 conditions. (A) Percent identification score of both vowels, (B) percent vowel segregation, (C) percent identification of one vowel of the pair, and (D) F_0 benefit comparison between data and model scores. Note that the legend mentioned in panel (A) also applies to panel (B) and the legend mentioned in panel (C) also applies to panel (D).

[see Fig. 1 for data or Fig. 3(A) for model response] are largely influenced by the level-dependent changes in phase locking of AN fibers to second-vowel characteristics (i.e., either formants or F_0 , or may be both). Figure 3(D) shows the actual (dashed line) and predicted F_0 benefit (solid line) at each vowel level. Despite the model's F_0 benefit being lower across levels, the model could successfully capture the pattern of variation in actual F_0 benefit with vowel level, qualitatively. The lowest F_0 benefit at 25 dB SPL in the data or model [Fig. 3(D)] could be associated with the limited F_0 -based segregation [see Fig. 3(B)] and thus confirms one of the suggestions from Chintanpalli *et al.* (2014).

The model's identification score was deterministic, even with the multiple repetitions of same concurrent vowel. However, the listeners' responses include response variability that was not captured in the current modeling. To compensate this effect to some extent, the similarity scores were used to determine how well the model captures either the listeners' correct answer or confusion. A similar type of this metric was also used by Chintanpalli and Heinz (2013) to evaluate their model performance in terms of listeners' responses at a single vowel level but here the current study further extends across vowel levels. As per this metric, for a given concurrent vowel, the model was given 1 point if it predicted the correct answer, 0.5 points if it predicted the first major confusion, 0.25 points if it predicted the second confusion, and 0 points if it predicted the response that was not observed in listeners' responses. Figure 4 shows the percent similarity score, across all pairs, as a function of vowel

level for both same- (squares) and different- F_0 (triangles) conditions. For the same- F_0 condition, similarity scores were lower across levels. This suggests the model had the most difficulty in accounting for listeners' responses when there was no F_0 -based vowel segregation. Thus, the modified version of Meddis and Hewitt (1992) segregation algorithm can account for the level-dependent changes in percent correct identification scores for same- and different- F_0 conditions [Fig. 3(A)], but it does not fully account for concurrent-vowel identification data by listeners when specific confusions are considered.

To understand the availabilities of formant and F_0 difference cues underlying the level-dependent changes in concurrent vowel identification, a specific vowel pair /æ, a/ was analyzed for illustration purposes. Both of these vowels have same F_1 ($= 750$ Hz) and they differ with each other in terms of F_2 and F_3 (see Table I). Figure 5 shows the model response for /æ ($F_0 = 100$ Hz), a ($F_0 = 100$ Hz)/ presented at 25 dB SPL, 65 dB SPL, and 85 dB SPL (see individual columns). As expected, there was no F_0 -based segregation at each level and hence the model response was made from a single PACF using the m_1/m_2 criterion (see step 2 of Fig. 2). Figures 5(A)–5(C) show the individual ACF channels with CFs, logarithmically spaced between 125 Hz and 4000 Hz, for 25 dB SPL, 65 dB SPL, and 85 dB SPL, respectively. The estimated F_0 was correctly identified as 100 Hz, which was indicated by an arrow in PACF at each level [see the bottom panels; Figs. 5(D)–5(F)]. For vowel identification at each level, the timbre region of the PACF was compared with the

FIG. 4. Similarity scores as a function of vowel level for same and different F_0 conditions to evaluate the effectiveness of the model with respect to listeners' responses.

previously stored templates of five different single vowels. For 25 dB SPL, the model response was /æ, æ/, which was the second confusion (identified as 12%) in listeners' responses, but the correct answer was identified as 44%. The

same confusion was predicted by the model for 65 dB SPL; however, it was reduced to 7% and the correct identification score was increased to 56%. The model response was correct for 85 dB SPL but it was identified as only 28% in listeners' responses. The first confusion in this case was /æ, æ/ (identified as 29%) in the concurrent-vowel data and the model can successfully predict this confusion as well if the model was forced to pick only the identical vowels by increasing the m_1/m_2 value. With m_1/m_2 being fixed across levels, the model responses were successful in either predicting the confusion or correct answer across levels. However, only for 65 dB SPL, if m_1/m_2 was changed to 2.505, then the model response was correct. If the model response had to be correct only for 25 dB SPL, then m_1/m_2 should be very high ($>=5.9$). All these findings suggest that the model responses may be influenced by the level-dependent changes in cochlear non linearities and m_1/m_2 value of the Meddis and Hewitt (1992) algorithm.

Figure 6 shows the model responses for /æ ($F_0 = 100$ Hz), α ($F_0 = 126$ Hz)/ presented at 25 dB SPL. The peak in the PACF occurred at 10 ms, which corresponds to correct F_0 estimation (= 100 Hz), as indicated by an arrow in PACF [Fig. 6(B)]. The percent ACF channels that showed a peak at 10 ms was 44%, which was marginally higher than the user-

FIG. 5. Model responses for /æ ($F_0 = 100$ Hz), α ($F_0 = 100$ Hz)/ presented at 25 dB SPL (first column), 65 dB SPL (second column), and 85 dB SPL (third column). For panels (A)–(C), the individual ACF channels are computed from 100 different AN fibers, ranged logarithmically between 125 Hz and 4000 Hz. These channels are added together to obtain the pooled ACF [bottom panels; (D), (E), and (F)]. The estimated F_0 is 100 Hz, as indicated by the arrow. The model responses are shown in the bottom panels. The timbre regions of the /æ/ (solid line) are shown in the bottom panels and /α/ template (dotted line) is shown only in panel (F). Note that these templates are shown with an arbitrary vertical and horizontal offset for clarity. For visualization purposes, only 50% channels are shown in the ACF plots.

FIG. 6. Model responses for /æ ($F_0 = 100$ Hz), α ($F_0 = 126$ Hz)/ presented at 25 dB SPL. The caption is similar to one of the columns in Fig. 5.

defined segregation parameter (43%). Thus, the model identification is based only using the $m1/m2$ criterion. The model response was /æ, æ/, which was the first confusion (identified as 31%) in listeners' responses. Additionally, if $m1/m2$ ratio was changed to 2.56, then the model was successful in predicting the correct answer.

Figure 7 shows the model responses for /æ ($F_0 = 100$ Hz), α ($F_0 = 126$ Hz)/ presented at 65 dB SPL. There was a peak at 10 ms in the pitch region of the PACF [Fig. 7(D)] and thus the dominant $F_0 = 100$ Hz was estimated correctly. The model does F_0 segregation as 50% of the ACF channels showed a peak at 10 ms, which was less than the segregation parameter (52%). The ACF channels that had a peak at 10 ms were grouped together [see Fig. 7(B)], whereas the remaining channels were grouped separately [see Fig. 7(C)]. Two segregated PACFs were computed from these groups. For vowel identification, the timbre region of the segregated PACF was compared with the previously stored templates of five different single vowels. The model identified /æ/ and /α/ correctly [see Figs. 7(E) and 7(F)]. The respective F_0 of the vowel was correctly identified only in one case [i.e., $F_0 = 100$ Hz, Fig. 7(E)] but in another case, the PACF indicated that there were two prominent peaks at 100 Hz, PACF amplitude = 1.74, and 126 Hz, with PACF amplitude = 1.49 [Fig. 7(F)]. Since only one of the F_0 s needs to be estimated correctly for vowel segregation using F_0 difference (Meddis

and Hewitt, 1992), the model was successful in identifying both the vowels.

Figure 8 shows the model responses for /æ ($F_0 = 100$ Hz), α ($F_0 = 126$ Hz)/ presented at 85 dB SPL. The dominant F_0 was estimated correctly, as there was a peak at 7.9 ms (marginally higher than 10 ms) in the pitch region of the PACF [Fig. 8(D)]. Since only 24% of the ACF channels had a peak at 7.9 ms, F_0 based vowel segregation was allowed prior to its identification. The model identified /α/ with $F_0 = 126$ Hz [Fig. 8(E)] correctly but incorrectly responded to /α/ with $F_0 = 100$ Hz [Fig. 8(F); note that the second peak in the PACF corresponds to 126 Hz]. The response of the model was /α, α/, which was the second confusion (identified as 11%) and the correct identification score was 55% observed in listeners' responses. At 85 dB SPL, due to improper vowel segregation, the model failed to obtain the correct answer.

IV. DISCUSSION

A. Sensitivity of predicted concurrent-vowel identification scores to AN-model parameters

One recurrent question that arises in studies that use AN models (largely derived from animal studies) to predict human performance is whether the exact parameters used for cochlear frequency selectivity and SR distributions are appropriate for humans. Thus, this section investigates how changes in the AN-model parameters [e.g., tuning and SR] affect identification score for both vowels. There is a strong evidence that human tuning is sharper than animals (e.g., Shera *et al.*, 2002; Joris *et al.*, 2011), and thus we have directly altered the auditory-filter bandwidth at each CF to simulate sharper tuning in the cat AN model (Zilany *et al.*, 2014). To address the sensitivity issue associated with changes in SR, the default values for high, medium, and low-SR fibers (i.e., 100, 4, and 0.1 spikes/s) of the AN model (Zilany *et al.*, 2014) were altered to 20, 2, and 0.01 spikes/s, respectively. For both conditions (i.e., changes in tuning and SRs), it was found that the model scores still can successfully captured the qualitative level-dependent effects for the same- and different- F_0 conditions. The main requirement was that the pattern of level-dependent changes in vowel segregation for the different- F_0 condition has to be preserved and also that the vowel segregation was zero at each level for the same- F_0 condition [Fig. 3(B)]. More specifically, the pattern of vowel segregation with increasing level for the different- F_0 condition has to be similar, although the absolute values could differ. Thus, the overall conclusions of this modeling study do not appear to be dependent on the specific parameter choices related to frequency selectivity and spontaneous-rate distributions. It is possible that a better quantitative fit to the concurrent-vowel data could be obtained by altering the $m1/m2$ value slightly.

B. Effect of $m1/m2$ on identification of identical vs different vowels

The selection of $m1/m2$ will influence the model response (either identical or different vowels; see Fig. 2)

/æ, α/ presented at 65 dB SPL
Different F0 condition

FIG. 7. Model responses for /æ ($F_0 = 100$ Hz), α ($F_0 = 126$ Hz) presented at 65 dB SPL. The first column corresponds to the individual ACF channels from 100 different AN fibers. These channels are added together to obtain the pooled ACF [panel (D)]. The two fundamental frequencies in the panel (D), are denoted by F_{01} and F_{02} . The estimated F_0 is F_{01} ($= 100$ Hz). The second column shows only ACF channels that have peak at 10 ms [panel (B)] and the remaining channels are placed in the third column as shown in panel (C). The model responses are correct as shown in panels (E) and (F). Note that the timbre regions of the templates /æ/ and /α/ (solid lines) are shown in panels (E) and (F) with an arbitrary vertical and horizontal offset for clarity. For visualization purposes, only 50% channels are shown in the ACF plots.

under the no- F_0 -segregation condition and will greatly affect the percent correct identification. Thus, in the current study, m_1/m_2 value was varied from 1.25 to 3, with increments of 0.25. It was found that $m_1/m_2 = 2.5$ was successful in predicting the identification scores of both vowels [Fig. 3(A)] and F_0 benefit [Fig. 3(D)] at least qualitatively, and one-vowel identification quantitatively [Fig. 3(C)]. Additionally, when averaged across vowel pairs and levels for the same- F_0 condition, the percentage of model responses that were either correct or were a confusion observed in the data was 78%. For the different- F_0 conditions, the model had only 45 vowel pairs (out of 150) across levels that went to a no- F_0 -segregation condition and the percent model response either for correct answer or observed confusion was 82%.

C. Selection of single-vowel templates for identification

The optimal single-vowel templates at each level (i.e., five vowel templates per level) were used for identifying the vowels. The model response could be influenced by the manner in which the single-vowel templates are generated. For example, when a single template for each vowel was generated regardless of level (i.e., only five vowel templates), it

was found that model scores did not capture the patterns of level-dependent changes in identification scores for the same- and different- F_0 conditions (results not shown here). Thus, this analysis further suggests that the templates have to be different for each level due to the level-dependent changes in cochlear nonlinearities (e.g., Ruggero *et al.*, 1997).

D. Possible physiological mechanisms underlying the level-dependent changes in identification scores of concurrent vowels

The model scores of concurrent-vowel identification suggest that the level-dependent changes in phase locking of AN fibers to vowel formants and F_0 s can account for trends in segregation and identification of concurrent vowels across levels and F_0 conditions. Thus, the present modeling scores provide quantitative support for the qualitative suggestions from Chintanpalli *et al.* (2014). However, the model responses were lower across levels and F_0 difference, as compared to identification data [compare Fig. 3(A) vs Fig. 1], suggesting that other cues, apart from phase locking, could also be contributing. Rate-place cues could further improve the scores for low-to-mid levels, as it has been

FIG. 8. Model responses for /æ ($F_0 = 100$ Hz), a ($F_0 = 126$ Hz)/ presented at 85 dB SPL. The figure caption is similar to Fig. 7. The estimated $F_{01} = 126$ Hz. The second column shows only ACF channels that have peak at 7.9 ms [panel (B)] and the remaining channels are placed in the third column as shown in panel (C). The model responses are shown in panels (E) and (F). The model response is correct for panel (E) but incorrect for panel (F). The timbre region of the template /a/ (solid line) is shown in the bottom panels of (E) and (F).

shown to encode vowel's formants (e.g., [Sachs and Young, 1979](#)) and F_0 s ([Keilson et al., 1997](#)).

The effect of medial olivocochlear reflex (MOCR) can enhance the AN response to tone in the presence of background noise (e.g., [Kawase et al., 1993](#); [Chintanpalli et al., 2012](#); [Smalt et al., 2014](#)). However, this MOCR effect may not be contributing because the concurrent vowels require the identification of both vowels. Other central auditory cues (e.g., cognition, attention) can also interact with these identification scores of both vowels. However, the modeling adopted in this study did not take into account the rate-place of the periphery and central auditory cues.

E. Conclusion and Future Work

The present computational modeling study predicted the level-dependent changes in identification scores of both vowels for the cases with and without F_0 difference between the vowels being presented. The model was developed by cascading the physiologically realistic AN model with a modified version of Meddis and Hewitt F_0 -based segregation algorithm (for complete block diagram, see Fig. 2). The model was tested against the identification score of both vowels and identification score of one vowel of the pair as a function of vowel level for same- and different- F_0 conditions. The segregation parameter of [Meddis and Hewitt \(1992\)](#) modeling was controlled by the actual F_0 benefit

across levels. The results from the modeling revealed that vowel segregation (either using the F_0 difference or formant difference) based on the temporal coding cues of AN fibers can qualitatively explain the level-dependent changes in identification scores of both vowels with and without F_0 difference [Fig. 3(A)]. Additionally, the model was successful in capturing the identification of one vowel quantitatively [Fig. 3(C)] and F_0 benefit qualitatively [Fig. 3(D)].

Possible future work could include (1) model scores may be able to improve at low-and-mid levels by incorporating the rate place cues of AN fibers (e.g., [Sachs and Young, 1979](#)), in addition to the existing temporal cues. One way to include both the rate-place and temporal cues is to use a neural network framework for vowel identification. Perhaps neural networks need to be cascaded together such that the first network may provide F_0 segregation information, whereas the second network may provide identification scores. The input to the first neural network can be either temporal or rate-place cues or both. (2) The similarity score (Fig. 4) suggested that the model had greater difficulty in accounting for listeners' responses when there was no F_0 difference. This suggests that there is a need to improve the model performance when only formant difference cues are available (i.e., no F_0 difference). (3) The model can be extended to include response variability, as observed in listeners' responses. (4) Extensions to study the predicted

effects of sensorineural hearing loss on level-dependent changes in concurrent-vowel identification, using the same segregation parameters used in this study. These predictions can help us to design an appropriate experimental paradigm (e.g., suitable sound level) for hearing-impaired listeners.

ACKNOWLEDGMENTS

This work was supported by second author's Outstanding Potential for Excellence in Research and Academics (OPERA) Grant and Research Initiation Grant (RIG), awarded by BITS Pilani, Pilani campus, Rajasthan, India. The level-dependent concurrent vowel data were used from Chintanpalli *et al.* (2014), which were collected at Medical University of South Carolina, USA under the supervision of Dr. Judy R. Dubno (NIH/NIDCD: R01 DC000184 and P50 DC000422, and NIH/NCRR: UL1 RR029882), Department of Otolaryngology-Head and Neck Surgery. Many thanks to Dr. Michael G. Heinz for providing the valuable comments in constraining the model parameters and on the previous versions of this manuscript.

- Arehart, K. H., Arriaga, C., Kelly, K., and Mclean-Mudgett, S. (1997). "Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss," *J. Speech, Lang. Hear. Res.* **40**, 1434–1444.
- Arehart, K. H., Rossi-Katz, J., and Swensson-Prutsmann, J. (2005). "Double-vowel perception in listeners with cochlear hearing loss: Differences in fundamental frequency, ear of presentation, and relative amplitude," *J. Speech, Lang. Hear. Res.* **48**, 236–252.
- Assmann, P. F., and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," *J. Acoust. Soc. Am.* **88**, 680–697.
- Assmann, P. F., and Summerfield, Q. (1994). "The contribution of wave-form interactions to the perception of concurrent vowels," *J. Acoust. Soc. Am.* **95**, 471–484.
- Bernstein, J. G. W., and Oxenham, A. J. (2005). "An autocorrelation model with place dependence to account for the effect of harmonic number on fundamental frequency discrimination," *J. Acoust. Soc. Am.* **117**, 3816–3831.
- Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA), pp. 1–792.
- Brokx, J. P. L., and Nootboom, S. G. (1982). "Intonation and the perceptual separation of simultaneous voices," *J. Phon.* **10**, 23–36.
- Bruce, I. C., Sachs, M. B., and Young, E. D. (2003). "An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses," *J. Acoust. Soc. Am.* **113**, 369–388.
- Cariani, P. (2004). "A temporal model for pitch multiplicity and tonal consonance," in *Proceedings of the 8th International Conference on Music Perception and Cognition, SMPC*, August 3–7, Evanston, IL, pp. 310–313.
- Carney, L. H. (1993). "A model for the responses of low-frequency auditory-nerve fibers in cat," *J. Acoust. Soc. Am.* **93**, 401–417.
- Cherry, E. C. (1953). "Some experiments on the recognition of speech in one and both ears," *J. Acoust. Soc. Am.* **25**, 957–959.
- Chintanpalli, A., Ahlstrom, J. B., and Dubno, J. R. (2014). "Computational model predictions of cues for concurrent vowel identification," *J. Assoc. Res. Otolaryngol.* **15**, 823–837.
- Chintanpalli, A., Ahlstrom, J. B., and Dubno, J. R. (2016). "Effects of age and hearing loss on concurrent vowel identification," *J. Acoust. Soc. Am.* **140**, 4142–4153.
- Chintanpalli, A., and Heinz, M. G. (2013). "The use of confusion patterns to evaluate the neural basis for concurrent vowel identification," *J. Acoust. Soc. Am.* **134**, 2988–3000.
- Chintanpalli, A., Jennings, S. G., Heinz, M. G., and Strickland, E. A. (2012). "Modeling the anti-masking effects of the olivocochlear reflex in auditory nerve responses to tones in sustained noise," *J. Assoc. Res. Otolaryngol.* **13**, 219–235.
- Culling, J. F., and Darwin, C. J. (1993). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0," *J. Acoust. Soc. Am.* **93**, 3454–3467.
- Heinz, M. G., Zhang, X., Bruce, I. C., and Carney, L. H. (2001). "Auditory nerve model for predicting performance limits of normal and impaired listeners," *Acoust. Res. Lett. Online* **2**, 91–96.
- Joris, P. X., Bergevin, C., Kalluri, R., McLaughlin, M., Michelet, P., van der Heijden, M., and Shera, C. A. (2011). "Frequency selectivity in Old-World monkeys corroborates sharp cochlear tuning in humans," *Proc. Natl. Acad. Sci. U.S.A.* **108**, 17516–17520.
- Kawase, T., Delgutte, B., and Liberman, M. C. (1993). "Antimasking effects of the olivocochlear reflex II Enhancement of auditory-nerve response to masked tones," *J. Neurophysiol.* **70**, 2533–2549.
- Keilson, S. E., Richards, V. M., Wyman, B. T., and Young, E. D. (1997). "The representation of concurrent vowels in the cat anesthetized ventral cochlear nucleus: Evidence for a periodicity-tagged spectral representation," *J. Acoust. Soc. Am.* **102**, 1056–1071.
- Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* **67**, 971–995.
- Larsen, E., Cedolin, L., and Delgutte, B. (2008). "Pitch representations in the auditory nerve: Two concurrent complex tones," *J. Neurophysiol.* **100**, 1301–1319.
- Liberman, M. C. (1978). "Auditory-nerve response from cats raised in a low-noise chamber," *J. Acoust. Soc. Am.* **63**, 442–455.
- Meddis, R., and Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.* **91**, 233–245.
- Miller, R. L., Schilling, J. R., Franck, K. R., and Young, E. D. (1997). "Effects of acoustic trauma on the representation of the vowel /e/ in cat auditory nerve fibers," *J. Acoust. Soc. Am.* **101**, 3602–3616.
- Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. (1997). "Basilar-membrane responses to tones at the base of the chinchilla cochlea," *J. Acoust. Soc. Am.* **101**, 2151–2163.
- Sachs, M. B., and Young, E. D. (1979). "Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate," *J. Acoust. Soc. Am.* **66**, 470–479.
- Shera, C. A., Guinan, J. J., Jr., and Oxenham, A. J. (2002). "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3318–3323.
- Smalt, C. J., Heinz, M. G., and Strickland, E. A. (2014). "Modeling the time-varying and level-dependent effects of the medial olivocochlear reflex in auditory nerve responses," *J. Assoc. Res. Otolaryngol.* **15**, 159–173.
- Summerfield, Q., and Assmann, P. F. (1991). "Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony," *J. Acoust. Soc. Am.* **89**, 1364–1377.
- Summers, V., and Leek, M. (1998). "F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss," *J. Speech, Lang. Hear. Res.* **41**, 1294–1306.
- Tan, Q., and Carney, L. H. (2003). "A phenomenological model for the responses of auditory-nerve fibers. II. Nonlinear tuning with a frequency glide," *J. Acoust. Soc. Am.* **114**, 2007–2020.
- Vongpaisal, T., and Pichora-Fuller, M. K. (2007). "Effect of Age on F0 Difference Limen and Concurrent Vowel Identification," *J. Speech Lang. Hear. Res.* **50**, 1139.
- Young, E. D., and Sachs, M. B. (1979). "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Am.* **66**, 1381–1403.
- Zhang, X., Heinz, M. G., Bruce, I. C., and Carney, L. H. (2001). "A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression," *J. Acoust. Soc. Am.* **109**, 648–670.
- Zilany, M. S. A., and Bruce, I. C. (2006). "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," *J. Acoust. Soc. Am.* **120**, 1446–1466.
- Zilany, M. S. A., and Bruce, I. C. (2007). "Representation of the vowel /e/ in normal and impaired auditory nerve fibers: Model predictions of responses in cats," *J. Acoust. Soc. Am.* **122**, 402–417.
- Zilany, M. S. A., Bruce, I. C., and Carney, L. H. (2014). "Updated parameters and expanded simulation options for a model of the auditory periphery," *J. Acoust. Soc. Am.* **135**, 283–286.
- Zilany, M. S. A., Bruce, I. C., Nelson, P. C., and Carney, L. H. (2009). "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics," *J. Acoust. Soc. Am.* **126**, 2390–2412.
- Zwicker, U. T. (1984). "Auditory recognition of diotic and dichotic vowel pairs," *Speech Commun.* **3**, 265–277.