Speech Enhancement for Listeners with Hearing Loss Based on a Model for Vowel Coding in the Auditory Midbrain

Akshay Rao and Laurel H. Carney, Member, *IEEE*

A novel signal-processing strategy is proposed to enhance speech for listeners with hearing loss. The strategy focuses on improving vowel perception based on a recent hypothesis for vowel coding in the auditory system. Traditionally, studies of neural vowel encoding have focused on the representation of formants (peaks in vowel spectra) in the discharge patterns of the population of auditory-nerve (AN) fibers. A recent hypothesis focuses instead on vowel encoding in the auditory midbrain, and suggests a robust representation of formants. AN fiber discharge rates are characterized by pitch-related fluctuations having frequency-dependent modulation depths. Fibers tuned to frequencies near formants exhibit weaker pitch-related fluctuations than those tuned to frequencies between formants. Many auditory midbrain neurons show tuning to amplitude modulation frequency in addition to audio frequency. According to the auditory midbrain vowel encoding hypothesis, the response-map of a population of midbrain neurons tuned to modulations near voice-pitch exhibits minima near formant frequencies, due to the lack of strong pitch-related fluctuations at their inputs. This representation is robust over the range of noise conditions in which speech intelligibility is also robust for normal-hearing listeners. Based on this hypothesis, a vowel-enhancement strategy has been proposed that aims to restore vowel-encoding at the level of the auditory midbrain. The signal-processing consists of pitch tracking, formant-tracking and formant enhancement. The novel formant-tracking method proposed here estimates the first two formant frequencies by modeling characteristics of the auditory periphery, such as saturated discharge-rates of AN fibers and modulation tuning properties of auditory midbrain neurons. The formant enhancement stage aims to restore the representation of formants at the level of the midbrain by increasing the dominance of a single harmonic near each formant and saturating that frequency channel. A MATLAB implementation of the system with low computational complexity was developed. Objective tests of the formant-tracking subsystem on vowels suggest that the method generalizes well over a wide range of speakers and vowels.

***Index Terms*—Speech analysis, hearing aids, neural coding, formant detection, formant tracking, formant estimation, auditory models**

Received: January 16, 2014. Revised: March 9, 2014. This work was supported in part by NIH-DC010813.

A. Rao was with the University of Rochester, Department of Electrical and Computer Engineering, and is now at Bose Corporation, Framingham, MA 01701 (e-mail: Akshay_Rao@Bose.com).

L. H. Carney is with the University of Rochester, Departments of Biomedical Engineering and Neurobiology & Anatomy, Rochester, NY 14642 (e-mail: Laurel.Carney@Rochester.edu).

I. Introduction

Speech sounds are commonly classified into two major categories: vowels and consonants. Vowels are typically associated with higher energy and stronger periodicity. The relative importance of vowels and consonants in speech perception has been the topic of multiple studies. In studies using spoken sentences in the absence of background noise, vowels have been shown to play a more important role in word recognition than consonants [1-3]. In the presence of noise, vowels carry more speech information, possibly because formant cues are robust even in noise [4].

Formant frequencies correspond to peaks in the short-time energy spectra of voiced sounds, arising due to the resonances of the vocal tract. Formants are one of the major cues in vowel perception [5-7], along with other factors such as spectral shape [8, 9] and formant ratio [10, 11]. Multi-dimensional analysis of the perceptual vowel space has ascertained that the two dimensions that account for the most variance in the perceptual space correspond to the first two formant frequencies [12-14]. Investigations into the auditory neurophysiological bases of vowel perception [15] provide insight into how formant cues are encoded in the lower auditory system. Understanding vowel-encoding in the healthy and impaired auditory systems can help hearing-aid researchers identify specific problems.

Traditionally, vowel-encoding studies have focused on representations of formant cues in the output discharge-rates of auditory-nerve (AN) fibers [e.g., 16]. An AN fiber can be approximated as a band pass filter (Fig. 1(a)) tuned to a *characteristic frequency* (CF). Young and Sachs [16] and Delgutte and Kiang [17] showed that for conversational sound levels, discharge-rates of a population of AN fibers plotted as a function of CF are maximum at the formant frequencies of vowel stimuli. AN fiber responses synchronize to the fine structure of stimuli in the narrow band of frequencies around their CFs and also synchronize to envelope fluctuations. Voiced sounds have periodic spectra associated with a fundamental frequency (also known as voice pitch, or F0). Voiced sounds result in fiber discharge-rates that fluctuate at a frequency close to F0 [17], and the strengths of these fluctuations vary across the population of fibers. The discharge-rates of fibers tuned close to a formant exhibit weak fluctuations at F0 due to dominance of the harmonic closest to the spectral peak, referred to as synchrony capture (Fig. 1b,

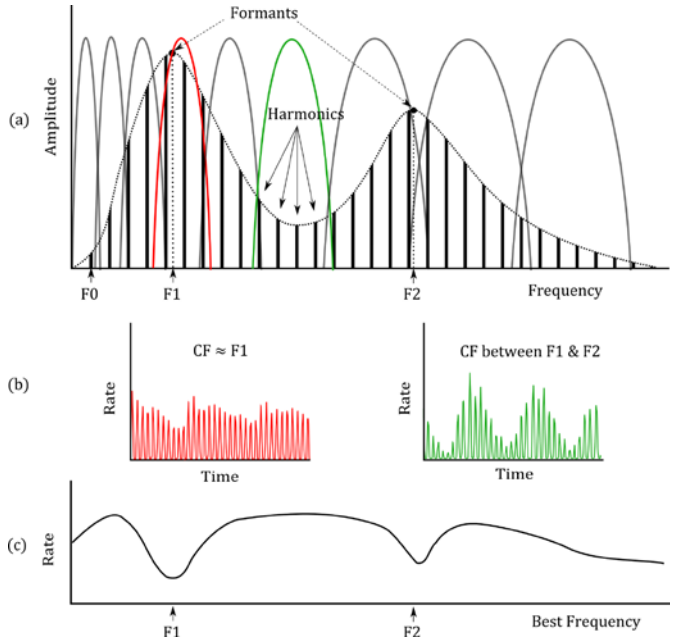

Fig. 1. (a) The simplified spectrum of a two-formant vowel is shown. Thick vertical lines are the harmonics of the fundamental frequency (F0). Spectral peaks at frequencies F1 and F2 are the formants for this vowel. Bandpass characteristics of a healthy auditory periphery are superimposed using gray, red and green lines. The bandwidths have been exaggerated for clarity. (b) The left panel shows the nearly-sustained discharge rates of an Auditory Nerve (AN) fiber with a characteristic frequency (CF) near F1 (in red). The right panel shows the discharge rates of an AN fiber with a CF between F1 and F2 (in green). The discharge rates in both panels fluctuate at the pitch period, but the left panel is more sustained due to the presence of a dominant harmonic at the input. The AN responses were simulated using the Zilany et al. (2009) AN model. (c) Schematic of the response of a population of auditory midbrain neurons tuned to modulations at F0. The response shows dips in the response of neurons having a best frequency near the formants.

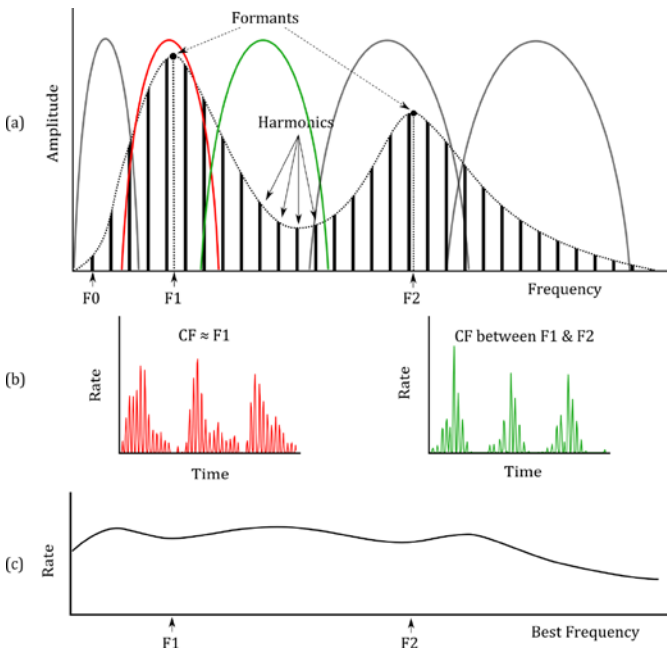

Fig 2. (a) Superimposed over the same two-formant vowel as Fig. 1, the bandpass characteristics of a highly impaired auditory periphery are shown. Note the broader bandwidths as compared to those in Fig. 1 due to hearing impairment. (b) The left panel shows the loss of the saturated characteristics of discharge rates of the AN fiber with a CF near F1 (in red) in an impaired auditory periphery, due to elevated thresholds. The right panel shows the discharge rates of an AN fiber with a CF between F1 and F2 (in green). The envelopes of both discharge rates are similar in depth. The AN responses were simulated using the Zilany et al. (2009) AN model. (c) A schematic of the response of a population of auditory midbrain neurons tuned to modulations at F0 in the impaired auditory system. The response no longer shows large dips in the response of neurons having best frequency near formants, indicating deterioration in vowel-coding. There is an overall loss of contrast between the response of neurons tuned near formants and those tuned between formants.

left panel). AN fibers tuned to frequencies intermediate to formants respond with strong fluctuations in their discharge-rates (Fig. 1b, right panel) due to "beats" between neighboring harmonics. Studies of vowel-encoding have traditionally focused on the synchrony of AN fiber discharge-rates to fine structure and on formant representation in the pattern of discharge-rates of a population of AN fibers. However, most proposed neural encoding strategies do not yet fully account for the robust nature of speech perception across a wide range of sound levels and in noise.

Studebaker, et al. [18] showed that speech recognition scores increase with sound pressure levels (SPL) above audibility thresholds and are highest at about 80 dB SPL, beyond which scores decrease. However, neural coding of formants based on the discharge-rates of a population of AN fibers deteriorates with increasing input sound levels [16]. Everyday communication exposes listeners to a large dynamic range of speech and noise levels, and thus it is important to understand the underlying cues that make neural coding robust in these conditions.

A recent vowel-coding hypothesis [15] focuses on neural coding of vowels at the level of the auditory midbrain. Many midbrain neurons are not only tuned to the energy within a narrow range around their best audio frequency or *best frequency* (BF), but are also tuned to the frequency of amplitude modulations [19-21]. That is, a midbrain neuron responds maximally to energy near its BF if the energy modulation rate is close to the neuron's *best modulation frequency* (BMF) (Fig. 3). Many modulation-tuned midbrain neurons in a wide range of species have BMFs between 10 and 300 Hz [reviewed in 22], which includes the range of voice pitch. According to the midbrain vowel-coding hypothesis, in addition to energy, the pitch-dependent strength of fluctuations in AN discharge-rates is significant in shaping midbrain neural responses. Also, as a consequence, a midbrain neuron with a BMF close to F0, exhibits lowered response rates if its BF is close to a formant and exhibits elevated response rates if its BF is between formants (Fig. 1c). The midbrain vowel-coding hypothesis is robust over a wide range of sound levels and tapers off for sound levels above 80 dB SPL. This neural coding strategy deteriorates for noise interference at signal-to-noise ratios consistent with listeners with normal hearing.

One of the underlying aims of any auditory neural encoding theory is to guide future hearing-aid research towards improving cues most important for normal hearing. It is thus also important to understand how a vowel-encoding scheme would account for the decrease in vowel discrimination in listeners with hearing loss. Sensorineural hearing loss is characterized in part by elevated thresholds and reduced

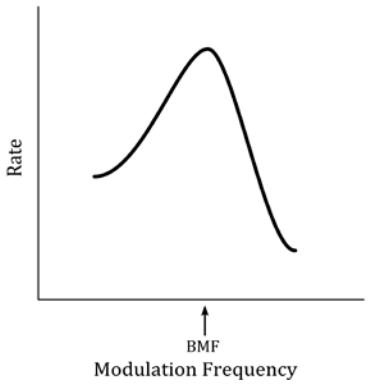

Fig. 3. Schematic of average-rate characteristics of a modulation-tuned auditory midbrain neuron. This figure shows the response to narrowband stimuli having energy close to the neuron's best audio frequency - or best frequency (BF) but modulated at different modulation frequencies. The modulation frequency at which the auditory midbrain neuron's response is maximum is its best modulation frequency (BMF).

frequency selectivity in the peripheral auditory system. To elicit fiber responses comparable to an un-impaired auditory periphery, higher input sound pressure levels are required. In addition, fibers respond to energy over a broader range of frequencies centered on their CF (Fig. 2a). A fiber with broader tuning and more linear response properties responds with stronger fluctuations in its discharge-rates, even when its CF is close to a formant, because the dominance of the harmonic closest to the spectral peak diminishes (Fig. 2b). Therefore, further up the auditory pathway, according to the midbrain vowel-coding hypothesis, midbrain neurons tuned near formant frequencies would show responses similar to neurons tuned between formants due to the strongly modulated envelope at their inputs. In other words, hearing loss results in a loss of contrast between the rate fluctuations of these neurons. Recall that, in an unimpaired auditory system, a midbrain neuron tuned to voice pitch exhibits a low response if its BF is near a formant. Consequently, without these "dips" in the midbrain neural response map, this encoding scheme predicts a reduction of formant information for listeners with hearing loss. Therefore, signal-processing strategies aimed at the restoration of formant information at the level of the auditory midbrain, can potentially lead to improvement in vowel discrimination in listeners with hearing loss.

The signal-processing strategy described here is based on the midbrain vowel-coding hypothesis. The strategy aims to improve vowel discrimination in listeners with hearing loss by restoring cues that are important for formant encoding at the level of the auditory midbrain. The signal-processing system tracks time-varying formants in voiced segments of the input and increases the dominance of a single harmonic near each formant in order to decrease F0-related fluctuations in that frequency channel.

II. Methods

This section describes the stages of the vowel-enhancement system. The general schematic of the vowel-enhancement system is shown in Fig. 4. The system consists of a speech analysis stage and a formant enhancement stage. The speech analysis stage performs various pre-processing tasks and

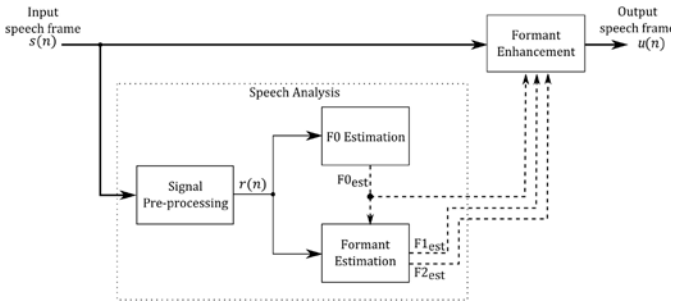

Fig. 4. Schematic of the vowel enhancement system. The system is divided into a Speech Analysis stage and a Formant Enhancement stage. The sub-stages of the Speech Analysis stage are also shown. Solid arrows indicate flow of the speech signal and dashed arrows indicate flow of calculated parameters such as pitch and formants.

estimates the fundamental frequency (F0) and the first two formants (F1 and F2) of the speech frame. The formant enhancement stage then amplifies the harmonic closest to each formant estimate, thereby increasing its dominance.

A. Signal Analysis

1) Signal Pre-processing: In the MATLAB-based implementation of the vowel-enhancement system, the incoming speech signal was divided into 32-ms long frames, with 50% overlap across successive frames. For the sampling rate of 8000 Hz, this translated into a frame length of 256 samples. First, DC offset removal was performed on the current frame, followed by windowing:

$$s_{\mathrm{zm}}(n) = s(n) - \bar{s}(n), for\ 0 \le n \le N-1, n \in \mathbb{Z} \quad (1)$$

$$w(n) = 0.5\left(1 - \cos\frac{2\pi n}{N-1}\right) \quad (2)$$

$$r(n) = s_{\mathrm{zm}}(n)w(n) \quad (3)$$

where $s(n)$ is a sequence representing the current input frame; n is an index that takes integer values between 0 and N-1; N is the frame length (in number of samples); $\bar{s}(n)$ is the mean of the sequence $s(n)$ over the frame; $s_{\mathrm{zm}}(n)$ is the zero-mean sequence obtained after DC removal; and $r(n)$ is the sequence obtained after windowing $s_{\mathrm{zm}}(n)$ using a Hanning window $w(n)$ of length N.

2) F0 Estimation: Voiced regions of speech (e.g., vowels) are associated with a pitch and a set of formants. The F0 estimation stage identifies the current frame as being either voiced or unvoiced, and estimates F0. Many F0 detection algorithms employ methods such as autocorrelation, average magnitude difference function, zero-crossing rates, etc., to estimate the principal period of a speech frame [reviewed in 23]. In the work presented here, a MATLAB implementation of an autocorrelation-based pitch extraction algorithm was used from the Speech and Audio Processing Toolbox [24].

Typical autocorrelation-based pitch extraction algorithms compute a running autocorrelation function (ACF) for each frame within a range of time delays. The frame's periodicity is indicated by the peaks in $c_{rr}(\delta)$ and the time delays (δ) corresponding to these peaks indicate the possible pitch periods (Fig. 5). The range of possible pitch periods was limited to 2.5 ms - 14.3 ms, corresponding to a plausible voice pitch range from 70 Hz to 400 Hz [25]. Another modification

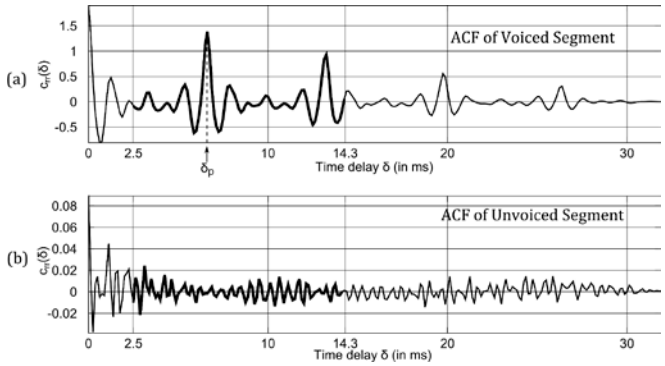

Fig. 5. Autocorrelation functions (ACF) of two 32 ms segments are shown. The horizontal axis represents the lag or time delay (δ) and the vertical axis represents the value of the ACF ($c_{rr}(\delta)$). The region of interest (between 2.5 ms and 14.3 ms) of the ACF is denoted by a thick curve. This region corresponds to the plausible range of voice pitch (70 Hz to 400 Hz). The highest peak from this region is considered the candidate pitch period. (a) ACF of a 32 ms segment of the vowel portion of the word 'had' is shown. The time lag corresponding to the maximum value of the ACF (δ_p) is the pitch period of this vowel (about 6.625 ms). It corresponds to a pitch (F0) of about 150.9 Hz. (b) ACF of a 32 ms segment of the leading consonant *h* of the word 'had' is shown. Note the differences in periodicity and scale of the vertical axis in its ACF as compared to that of the voiced segment.

was made to the ACF calculation in order to reduce the tapering off of the function due to decreasing overlap lengths at large values of δ. This tapering effect was reduced by using a variation of the ACF in which the sum is divided by the length of overlap ($N-\delta$):

$$c_{rr}(\delta) = \left(\sum_{n=0}^{N-1-\delta} r(n)\, r(n+\delta) \right) \Big/ (N-\delta) \qquad (4)$$

where $c_{rr}(\delta)$ is the autocorrelation sequence of the current frame $r(n)$; δ is the lag or delay (in samples); and N is the frame length (in samples).

The distinction between frames of interest (voiced frames) and silent or unvoiced frames was based on a *clarity* metric [26]. If for a particular frame, $c_{rr}(\delta)$ was found to be maximum at δ_p, then clarity of that frame was defined as the ratio $\frac{c_{rr}(\delta_p)}{c_{rr}(0)}$. High clarity indicates frames with voiced speech whereas low clarity indicates frames with unvoiced speech or silence. A frame's F0 estimate ($F0_{est}$) was set to 0 if its clarity was below a threshold. In the formant-tracking stage, frames with $F0_{est}$ equal to zero are considered to be unvoiced frames. A suitable threshold value for clarity for speech sentences in quiet was empirically found to be 0.50.

3) Formant Tracking: In this stage, the first two formants are estimated for the current voiced frame. Formant-tracking is not performed for frames with clarity below threshold. This stage replicates salient aspects of physiological auditory processing, such as the bandpass filtering of the auditory periphery, saturated discharge-rates of AN fibers, and the tuning of midbrain neurons to F0-related modulations. Sub-stages within the formant-tracking stage (Fig. 6) are described next.

a) *Auditory filtering:* The speech frame, $r(n)$, is decomposed into multiple bandpass channels, $x(f,n)$, by an auditory filterbank comprising a set of bandpass filters with center frequencies based on the equivalent rectangular bandwidth (ERB) scale [27]. An auditory filterbank reflects properties of the basilar membrane such as the logarithmic physical mapping of frequencies, and frequency-dependent bandwidths. These filterbanks consist of approximately logarithmically-spaced filters with bandwidths increasing with center frequency. The center frequencies of the 44-channel filterbank used here ranged from 70 Hz to 3700 Hz. The lower limit of this frequency range was chosen to match the lower limit of the plausible range of human voice pitch [25].

b) *Saturating non-linearity:* Each filter channel of the current frame is scaled on a sample-by-sample basis using a saturating nonlinearity. The nonlinearity serves to replicate the level-dependent discharge-rate saturation characteristics of AN fibers. Saturation is critical for the enhancement algorithm as it influences the degree of amplitude modulation within the channel. The sigmoid curve used was a Boltzmann function of the form:

$$x_{nl}(f,n) = \frac{A_1 - A_2}{1 + e^{x(f,n)\gamma(f)}} + A_2 \qquad (5)$$

where $x_{nl}(f,n)$ is the output of the nonlinearity for the bandpass-filtered channel $x(f,n)$ with center frequency f; A_1 and A_2 are the lower and upper limits of the nonlinearity and were fixed at -1 and 1 respectively; $\gamma(f)$ is the slope of the sigmoid curve and depends on the center frequency of the current channel. $\gamma(f)$ was determined using a frequency-dependent source spectrum threshold function based on a well-known model of speech production, described next.

According to the Source-Filter Model of Speech Production [5], speech sounds are the result of a source of sound energy (e.g., the larynx) and a vocal tract filter. The filter's transfer function is shaped by resonances of the vocal tract. In the case of voiced sounds (Fig. 7a), the magnitude spectrum of the sound source (known as source spectrum) contains peaks at F0 and at its harmonics, with a downward slope between 8 and 16 dB/octave [5, 28]. This monotonically decreasing source spectrum is then shaped by the transfer function of the vocal tract filter (Fig. 7b), resulting in the spectral peaks known as formants. Note that F0 is attenuated (Fig. 7(c)) by the vocal tract filter and is usually several dB less than the level at F1.

For a frame with index c, the slope of the nonlinearity ($\gamma_c(f)$) was calculated such that its output had an overall flat envelope for channels near formants, similar to the output discharge-rates of AN fibers tuned near formants. The source spectrum threshold function ($S_c(f)$) is a nonlinear function of frequency and decreases monotonically, similar to the peaks of the source spectrum in the source-filter model. $S_c(f)$ was defined as:

$$S_c(f) = \frac{10^{\frac{-m \cdot log_2(f/F0) - k}{20}}}{x_{rms}(F0)} \qquad (6)$$

where f is the center frequency of an auditory filter channel, and c is the index of the current frame; $F0$ is the voice pitch of

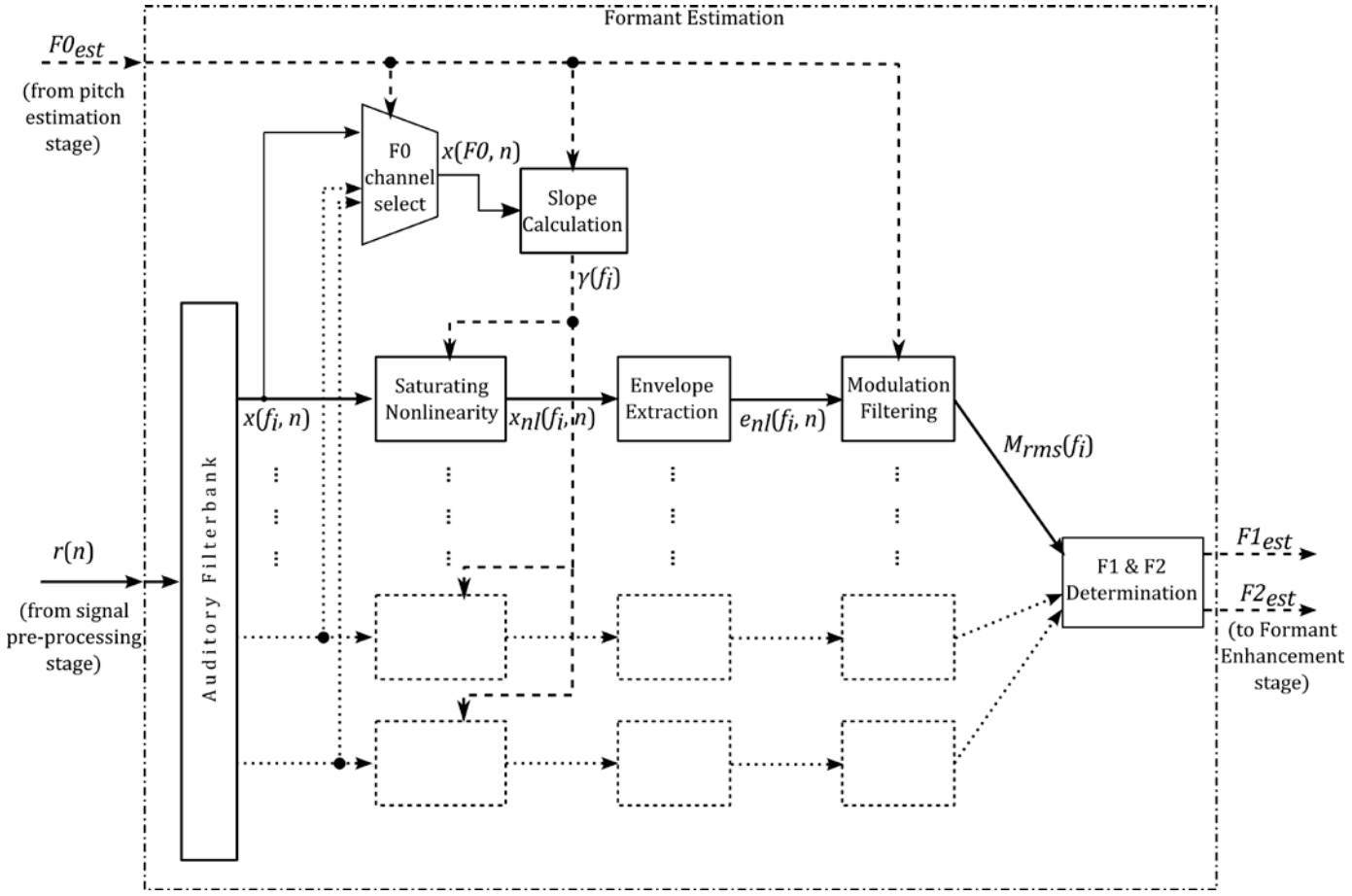

Fig. 6. Schematic of Formant Estimation. The inputs to this stage are the speech frame $r(n)$ from signal pre-processing and $F0_{est}$ from the pitch estimation stage. $r(n)$ is filtered into N bandpass channels using an auditory filterbank. $x(f_i, n)$ denotes the bandpass component of $r(n)$ centered at the frequency f_i. The signal chain for one bandpass channel at an arbitrary center frequency f_i is shown. Solid arrows represent flow of the speech signal, while dashed arrows represent flow of parameters such as pitch and formant estimates. Some, but not all, of the corresponding pathways for other channels have been shown using dotted arrows. For the sake of clarity, a few obvious signal paths and operations such as energy criterion and smoothing have been omitted from the schematic.

the current frame; $x_{rms}(F0)$ is the RMS value of the filter output whose center frequency is closest to $F0_{est}$ (denoted as F0 Channel Select in Fig. 6); m is the source spectrum slope (in dB/octave); and k (in dB) is a factor employed to partially offset the attenuation at F0 due to the vocal-tract filter. Suitable values of m and k were empirically determined (-9 dB/octave and 6 dB respectively) such that the RMS values of channels near formants remain above the source spectrum threshold value (Fig. 7(c)), and thus result in those channels being saturated to a higher degree by the sigmoid function than channels away from formants (Fig. 8).

The frequency-dependent slope $\gamma_c(f)$ of the nonlinearity was obtained using the following equation:

$$\gamma_c(f) = l \cdot S_c(f) \tag{7}$$

where l is a constant that controls the influence of $S_c(f)$ on the saturating nonlinearity. Decreasing l results in more aggressive saturation. In the current implementation, the value of l was set to 1.

c) *Envelope extraction:* In this stage, the envelope of each channel was obtained by removing the fine structure of the output of the nonlinearity ($x_{nl}(f,n)$) with a full-wave rectification followed by low-pass filtering with a cutoff frequency of 400 Hz with a 50th order FIR filter. The signal $e_{nl}(f,n)$ was then obtained by performing DC offset removal on the envelope of the signal. This was done in order to remove the influence of overall energy differences between channel envelopes before calculation of the pitch-related channel strengths in the next stage.

d) *Modulation filtering:* Next, modulation filtering was performed to simulate the modulation-tuning of auditory midbrain neurons. Each channel envelope was passed through a narrow bandpass filter centered at F0 to extract the signal components having frequency near F0. Then, in order to quantify the relative strengths of F0-related modulations across all channels, a measure $M_{rms}(f)$ was obtained by calculating the RMS of each channel envelope's F0 component. $M_{rms}(f)$ is thus a sequence indexed on the center frequency of each channel of the auditory filterbank. Due to the higher degree of saturation near formants, frequencies corresponding to the minima of $M_{rms}(f)$ were closest to the actual formants.

e) *F1/F2 determination:* Next, $M_{rms}(f)$ was smoothed using a 5-point symmetric, exponentially weighted smoothing kernel prior to locating its local minima. Center frequencies corresponding to the minima were selected as candidate formants and sorted in ascending order of frequency. In addition to saturation of channel outputs, minima in $M_{rms}(f)$

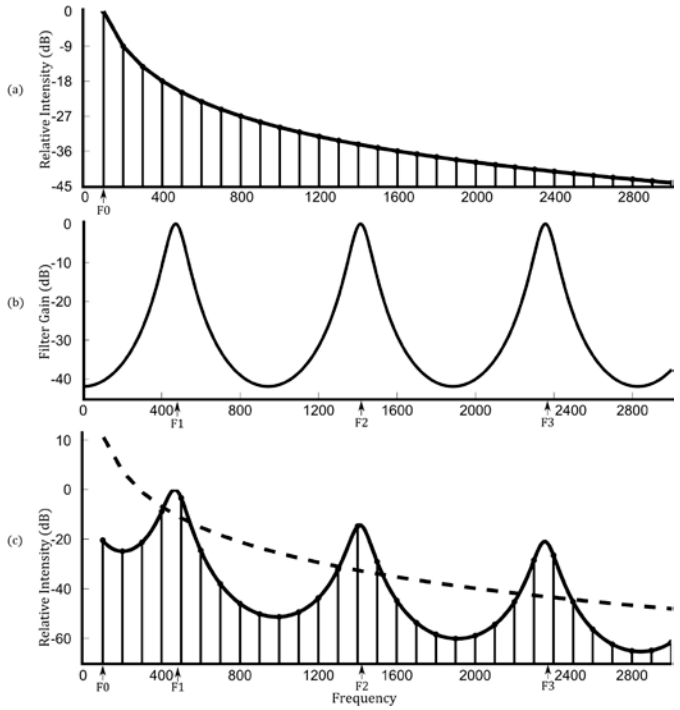

Fig. 7. (a) The spectrum of a sound source with F0=100 Hz is shown. Locations of the vertical lines represent harmonic frequencies and their length indicates the intensity of the harmonic. The source spectrum slopes (m) downwards at a rate of 9 dB/octave starting from F0. (b) The gain versus frequency plot of a vocal-tract filter with three spectral peaks is shown. (c) The spectrum of the resultant sound is shown. Notice that the intensity of the first harmonic (F0) is several dB less than that of the first formant. The dashed curve is an example source spectrum threshold function $S_c(f)$ calculated for an offset (k) of about 30 dB above the amplitude of F0. For clarity, figures and parameters are exaggerated here.

could also be due to very low energy in a particular channel. In order to eliminate such spurious minima, an energy criterion was imposed using the RMS values of the output of the saturating nonlinearity of each channel. A channel having an RMS value below the average of those RMS values was rejected as a possible formant channel. From the remaining values in $M_{rms}(f)$, the formant estimates $F1_{est}$ and $F2_{est}$ were obtained by choosing center frequencies corresponding to the first two values. The formant estimates were thus limited to center frequencies of auditory filters.

B. Formant Enhancement

This stage utilizes $F0_{est}$, $F1_{est}$ and $F2_{est}$ provided by the speech analysis stage to boost the dominance of a single harmonic near F1 and F2. According to the midbrain vowel-coding hypothesis, deterioration of formant-encoding at the level of auditory midbrain neurons can be attributed to broadened frequency selectivity properties of an impaired auditory periphery, resulting in a reduction in the dominance of the harmonic closest to formants. As a logical extension, artificially increasing the dominance of a harmonic was hypothesized to counter this phenomenon and lead to AN discharge characteristics more similar to those in the normal ear.

As shown in Fig. 9, first, the frequencies ν_1 and ν_2 of two harmonics were calculated by finding the integer multiples of $F0_{est}$ closest to $F1_{est}$ and $F2_{est}$. If any formant estimate was found to be equidistant from two adjacent harmonics, the lower harmonic was chosen.

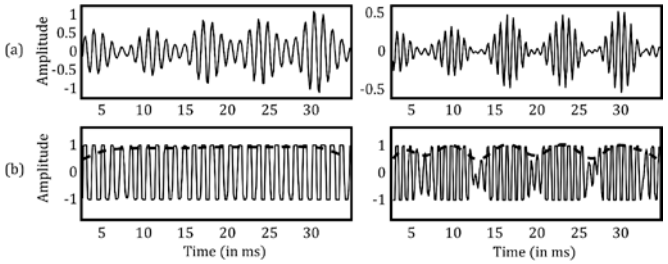

Fig. 8. (a) Waveforms of two bandpass channels are shown, one centered near a formant (left panels) and the other centered between formants (right panels). (b) The corresponding outputs of the saturating nonlinearity for both waveforms are shown. The output of the envelope extraction step is shown as thick dashed lines. As shown above, the nonlinearity saturates channels near formants to a higher degree than channels away from formants. Thus, the envelope (dashed lines) is much flatter for channels near formants.

Next, two linear-phase narrowband finite impulse response (FIR) bandpass filters, centered at ν_1 and ν_2 respectively, having passband gains of g_1 and g_2, amplified the respective harmonics in the current speech frame, $s(n)$. In the current implementation, an FIR filter of order 300 was generated using the Kaiser Window method of FIR filter design, using a bandwidth of 50 Hz and a stopband attenuation of 25 dB. A gain g_0 was then applied to the summation in order to account for elevated thresholds in listeners with hearing loss. Appropriate values of these gains would be determined empirically for each subject. The gains g_1 and g_2 would be fixed across time, and selected based on responses to a range of vowel sounds. Fig. 10 shows the spectrogram of a speech utterance before and after processing by the speech-enhancement system.

III. Results

A non real-time implementation of the system with tunable parameters was developed in MATLAB to test the ability of the vowel-coding hypothesis to guide a novel formant-tracking method and to enhance the discrimination of vowels in listeners with hearing loss.

The three parameters of the saturating non-linearity in the formant-tracking subsystem – k, l and m in (6) and (7), were deduced empirically using a speech dataset consisting of four

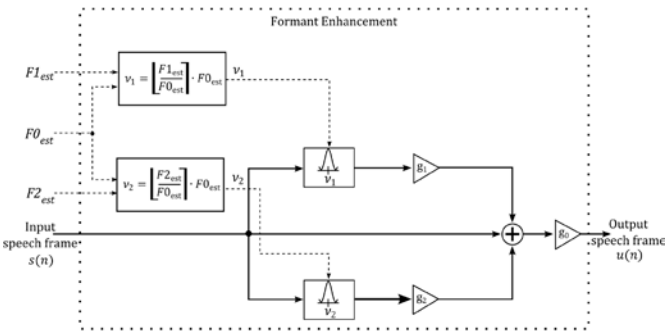

Fig. 9. Schematic of the formant enhancement stage. Pitch- and formant-estimates are used to first calculate the harmonics (ν_1 & ν_2) closest to the formants. The input speech frame is then filtered by two narrow bandpass filters centered at each calculated harmonic and gains g_1 and g_2 are applied respectively. A gain g_0 is applied to the summed signal, which results in the output speech frame. $\lfloor\cdot\rceil$ denotes rounding-off to the nearest integer.

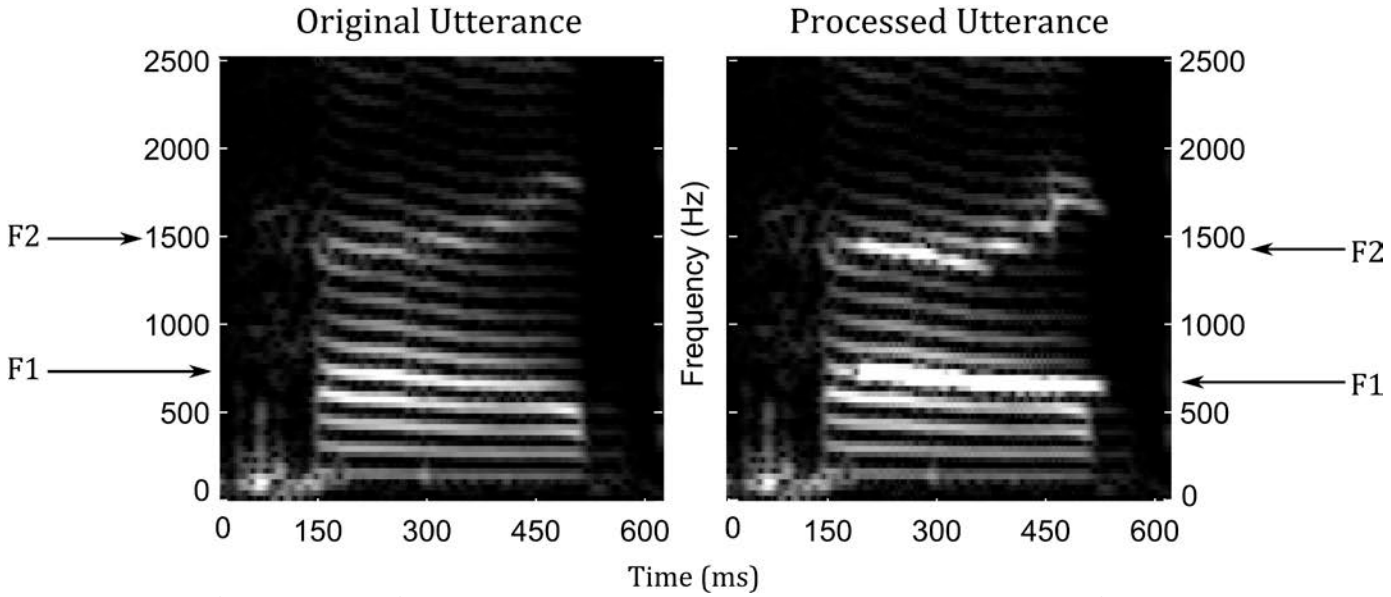

Fig. 10. Spectrograms of a single utterance of the word "hud" (/uh/) spoken by a male speaker and the corresponding output of the speech enhancement system are shown in the left and right panels, respectively. The vowel onset in the utterance is at about 150 ms. The z-axis (in grayscale) represents signal power (in decibels) with lighter colors representing regions of higher power. F1 begins approximately at 750 Hz while F2 begins at 1500 Hz. The trajectories of the first two formants are evident in the spectrograms with F1 decreasing in frequency with time and F2 increasing. The spectrogram of the output of the speech enhancement system (right panel) shows single harmonics closest to the formant estimates enhanced by gains of 15 dB (g_1) and 12 dB (g_2) respectively. The overall gain (g_0) was 0 dB. These gains were chosen for greater contrast between the two panels. The recording of this speech utterance was provided by Professor J. McDonough (Department of Linguistics, University of Rochester).

vowels: /ae/ ("had"), /iy/ ("heed"), /uw/ ("who'd") and /uh/ ("hud") from one male speaker. These recordings were provided by Professor J. McDonough (Department of Linguistics, University of Rochester). Keeping these parameters fixed, the formant-tracking subsystem was then evaluated using a vowel database containing 12 English vowels spoken by 139 speakers consisting of 93 adults (male and female speakers) and 46 children (27 boys and 19 girls). The database consists of single-vowel samples of the form "h*V*d", where *V* is an English vowel [30]. This annotated database contains acoustic measurements of each vowel sample including vowel durations, start and stop-times, and pitch and formant values at the middle of the vowel duration.

In order to compare estimates of the formant-tracking subsystem to the database formant values, the vowel portion from each sample was extracted using the vowel start and end times provided by the database. This segment was then downsampled to 8000 Hz and was passed through the pitch tracking and formant-tracking subsystems. Next, $F0_{est}$, $F1_{est}$ and $F2_{est}$ of the center-most frame were selected. The magnitude of the difference between each formant estimate and its corresponding known formant frequency from the database was normalized using the known F0 value. This measure of error gauges the deviation of the estimates in terms of number of harmonics, for example, values of this measure between 2 and -2 indicate that the formant estimate was correct within two harmonics. Vowel utterances for which the pitch tracking system wholly failed to identify the center-most frame as a voiced frame were not shown. Approximately 5.42% of the vowel utterances in the database were discarded for this reason.

Fig. 11 shows the performance of the formant-tracking subsystem on four vowels from the database that were matched to the vowels used to empirically determine the subsystem parameters. Each group of four data columns represents one vowel and each column within a group represents a particular speaker type. Each point on the scatter plot represents an individual vowel utterance. The vertical position of each point represents the normalized error or harmonic distance between a formant estimate and its true value. The shaded region represents errors within one harmonic of the true formant. The panels on the left (Fig. 11) show the full range of F1 and F2 estimation errors. The panels on the right zoom in on the region of the errors that lie within three harmonics for F1 estimation and eight harmonics for F2 estimation. Similar to Fig. 11, Fig. 12 shows the performance of the subsystem on the remaining 8 vowels from the database.

Figs. 11 and 12 show that a majority of the errors are within two harmonics for F1 estimates and within five harmonics for F2 estimates. The results indicate that the formant-tracking subsystem generalizes over multiple speakers and vowels reasonably well.

DISCUSSION

A novel physiologically-based signal-processing strategy for vowel enhancement and formant-tracking was developed. Targeting formants for improving speech perception has been proposed in previous studies, for example, the Contrast-enhancing frequency shaping method, aimed at producing better representation of formants at the output of the auditory-nerve fiber responses [31]. This method used a time-varying high pass filter to amplify the region of the vowel-spectrum above F2. Extensions of this method amplified formants above and including F2 without applying gains between formants. These methods, based on traditional vowel-encoding theories,

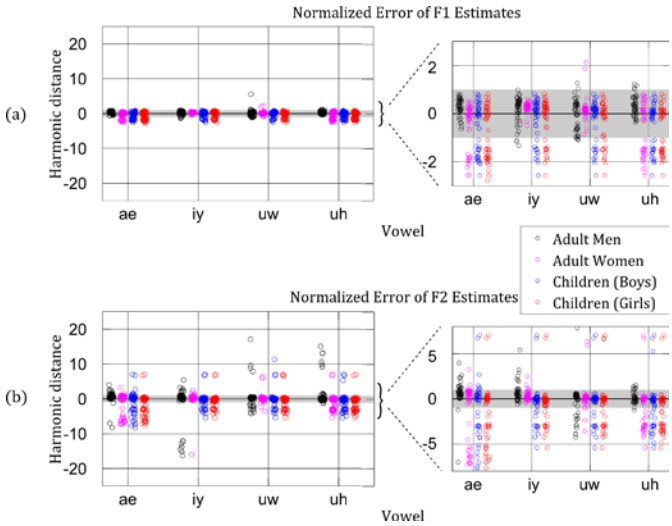

Fig. 11. This figure shows the performance of the formant tracking subsystem on four vowels (/ae/, /iy/, /uw/ and /uh/) spoken by 45 adult men, 48 adult women, 27 boys and 19 girls. (a) Performance of F1 estimation. Each data point represents the estimation error for an individual utterance of a vowel. Each group of four columns of symbols is a set of utterances of the same vowel spoken by the four types of speakers. The vertical position of each point indicates the harmonic distance between $F1_{est}$ and (known) F1. The left panel shows the full range of estimation errors while the right panel zooms in on the region with errors within three harmonics. The shaded region shows the errors lying within one harmonic. (b) Performance of F2 estimation. The left panel shows the full range of estimation errors and the right panel limits the range to within eight harmonics.

centered on energy in speech stimuli near formants and were guided by response characteristics of AN fibers. The midbrain vowel-encoding hypothesis focuses on the envelope-coding properties of auditory midbrain cells. In this strategy, narrowband filters were employed to boost single harmonics near the first two formants. This scheme focuses on the restoration of amplitude-modulation characteristics in the responses of auditory-nerve fibers. In particular, the strategy weakens pitch-related fluctuations of AN fiber discharge-rates for frequency channels near formants, and by extension, aims to restore the contrast in modulation characteristics across the population of frequency channels that provide the inputs to auditory midbrain neurons. The strategy is focused on vowel sounds, but the approach applies to any voiced sound, which includes some consonant sounds. Whether enhancement of the voiced sounds can compensate for the decreased intelligibility of consonants in a noisy background will be tested in future studies.

The formant-tracking method guided by physiology presents a novel approach to the problem of formant estimation. In the past, many schemes have been proposed for formant-tracking. In some parametric approaches [e.g., 32, 33, 34], all-pole or pole-zero linear prediction models were employed to perform spectral fitting of short-time vowel spectra. Non-parametric methods based on observations of local energy maxima and frequency modulation near formants can also be found in the literature [e.g., 35, 36, 37]. A physiologically based formant-tracking subsystem proposed by Delgutte [38] was based on energy-related response characteristics of auditory-nerve fibers and analyzed response patterns of filters of an auditory filterbank to estimate formant frequencies using the

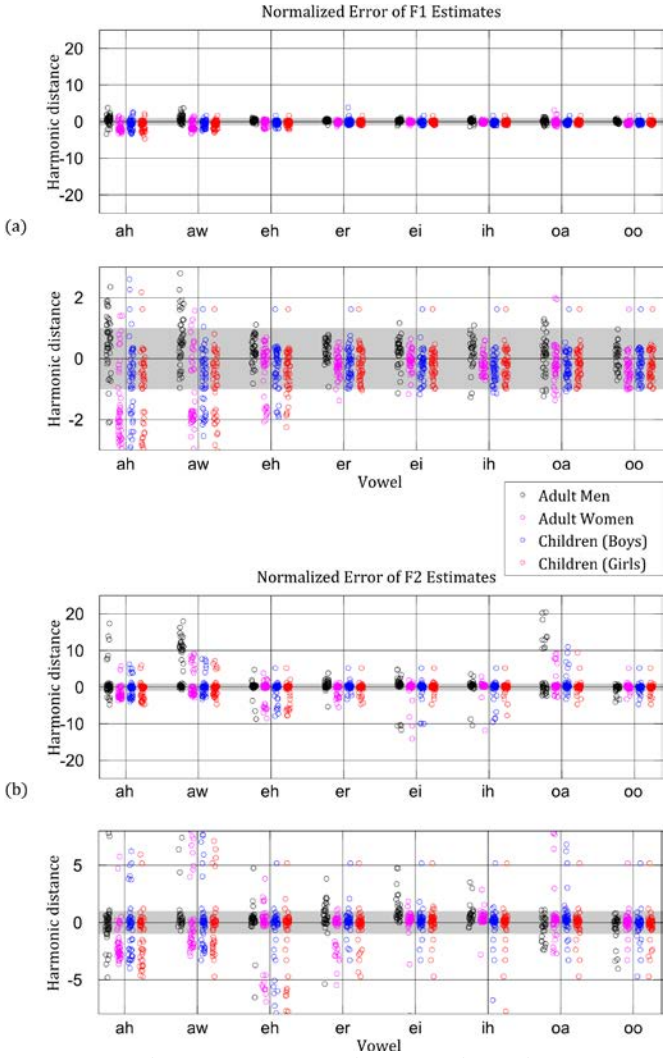

Fig. 12. This figure shows the performance of the formant-tracking subsystem on eight vowels. A separate set of four vowels was used to determine the parameters of the formant-tracking subsystem (see Fig. 11). (a) and (b) show the performance of F1 and F2 estimation respectively. The top panels show the full range of estimation errors while the bottom panels focus on the errors lying within three harmonics for F1 estimation and eight harmonics for F2 estimation.

distribution of energy across channels. The formant-tracking method described here is unique in its physiological basis on observed amplitude-modulation characteristics of AN fiber responses and the modulation-tuning properties in auditory midbrain neurons. Evaluation of the formant-tracking subsystem as well as the vowel-enhancement system will be discussed next.

A. Evaluation of Formant-tracking and Vowel Enhancement Strategies

Objective evaluation experiments were carried out to test the formant-tracking subsystem on single vowels on a large number of speakers. The vowel database contains speech samples from 139 American English speakers speaking 12 vowels, and was therefore used to evaluate the ability of the formant-tracking subsystem to generalize over different speakers and vowels. The results of the objective tests (Figs. 11, 12) show that the formant-tracking strategy is likely to generalize well over multiple speakers. The algorithm

performed more poorly for F2 estimates than for F1 estimates, and this trend was seen across speaker types and vowels. The majority of F1 estimation errors are below one harmonic, whereas they are below five harmonics for F2 (Figs. 11, 12). The possible reasons for this trend and general estimation errors are discussed in the next sub-section.

Comparison of results for all 12 vowels indicates that the formant-tracking strategy generalizes well over many vowels, including those not used for determination of the system parameters. Further fine tuning of the system parameters may be required in order to achieve higher accuracy for F2 estimates. Many formant-tracking techniques in the literature include gender detection modules to apply different processing or different parameters for male and female speakers. However, the performance of the formant tracking subsystem yielded similar results for adult speakers of both genders, in addition to children. Objective evaluation tests for vowels spoken in noise would reveal the suitability of this strategy to real-world sounds.

B. Sources of Errors, Issues and Improvements

In addition to frequencies corresponding to channels close to formants and those with low energy, minima were also found in channels in the neighborhood of those close to formants. Many of these minima occur at channels corresponding to the first few harmonics of the speech sample and are more defined for speakers with high voice pitch (women and children). These contribute to most of the F1 estimation errors and some of the F2 under-estimation errors where a minima at a frequency close to F1 (but higher in frequency) is selected as F2. Problems due to these smaller minima can be reduced with a more aggressive smoothing function and better minima-calculation techniques.

Vowels having F1 and F2 close to each other (e.g. /aw/) are more prone to F1 and F2 overestimation errors due to the merging of F1/F2 minima due to smoothing. In these cases, F3 is misidentified as F2. This case, combined with multiple minima near formants presents the tradeoff that the smoothing operation needs to overcome. Aggressive smoothing reduces overall F1/F2 estimation errors but would result in insufficient separation of F1/F2 minima in vowels with close F1/F2 frequencies. Another factor for F1/F2 estimation errors is that in some cases, $M_{\mathrm{rms}}(f)$ exhibits broad and flat regions of minima with multiple undulations within the region. This causes one of those undulations to be misidentified as a formant.

$M_{\mathrm{rms}}(f)$ was smoothed before minima calculation (section A.3.e in Methods). Smoothing is essential for minima calculation because $M_{\mathrm{rms}}(f)$ may contain similar values at points adjacent to the center frequency corresponding to a formant. Therefore, during preliminary testing on a single utterance of four vowels, the formant estimate was erroneously calculated by the system as the center frequency adjacent to the one closest to the true formant. In a few cases, this "bias" was found to be towards higher center frequency; however this does not seem to be the general case as Figs. 11 and 12 show a bias towards underestimation by the formant-tracking subsystem. It is speculated that the role of smoothing may be a contributing factor for this phenomenon due to logarithmic spacing between each center frequency. Instead of symmetric smoothing weights, asymmetric weights may be required to account for the unequal distance between successive center frequencies. Asymmetric exponential smoothing weights were found to improve this problem in a few initial test cases, but a set of weights that generalized well could not be found trivially. For minima calculation, simple derivative-based minima techniques fail to apply due to the small number of points (one point for each center frequency of the auditory filterbank) and due to unequal spacing of the independent variable (center frequency). Minima calculation in the implementation was done using a built-in MATLAB function (*findpeaks*). To reduce errors due to low harmonics causing minima, the strongest minima is chosen from those that are within a spectral distance of 1.5 times the value of $F0_{est}$ from each other.

Formant estimation in vowels with low F1 frequencies (e.g. /ee/) can show large F1 over-estimation errors due to the effect of the slope offset factor (k in (6)) on low-frequency channels. When a formant is close to the pitch, the source spectrum threshold function is likely to remain higher than the energy at F1 because the difference in energy between F0 and F1 is lower than k. This leads to insufficient saturation of channels near F1 and thus, in those cases, the formant is likely to be ignored by the algorithm.

The pitch extraction subsystem is crucial for the performance of the formant-tracking subsystem and the overall vowel-enhancement system. The accuracy of $F0_{est}$ is important for the saturating nonlinearity's operation due to the dependence of its source spectrum threshold function (6) on the energy near the voice pitch. Additionally, the formant-tracking subsystem directly uses the distribution of the strength of F0-related fluctuations at the output of the modulation filter across each channel in the formant-tracking subsystem, which underscores the importance for accurate F0 estimation. A drawback of the simple variant of the autocorrelation function used (4) is that peaks corresponding to an integer multiple of the true pitch period may sometimes be the local maximum, resulting in $F0_{est}$ erroneously being calculated as half of the true pitch. This problem (called "pitch-halving") is common in computationally simple pitch extraction algorithms and can be reduced by either preserving the tapering effect observed in basic autocorrelation function, or more robustly, by detecting these errors through additional logic in the pitch extraction algorithm [23].

Another major purpose of the pitch extraction subsystem was to identify voiced regions of continuous speech because the operations of formant-tracking and vowel-enhancement are carried out on only the voiced portions of speech. Detection of voiced speech in the current implementation is done on the basis of a measure called clarity – the relative strength of the autocorrelation function at the delay corresponding to the candidate pitch period to its value at zero delay. Frames having high values of this ratio were deemed to be voiced. A simple binary decision like clarity is, however,

unable to fully generalize on a large range of real-world speech. These problems were also observed during preparation of preliminary test datasets consisting of English sentences spoken in quiet. Inaccuracies in voiced segment identification of some sentences were found and could be corrected by adjusting the clarity threshold of the pitch extraction algorithm. For robust pitch estimation and voiced region detection, other more reliable methods that satisfy computational constraints can be used instead [39-41].

The primary role of the saturating nonlinearity in the formant-tracking subsystem is to exaggerate the difference in depth of amplitude modulation between filter channels. Thus, analogous to the outputs of modulation-tuned auditory midbrain neurons, simple modulation filtering of channel outputs results in low RMS values of channels near formants. Objective evaluation tests have shown that the operation of the nonlinearity is robust over multiple speakers and vowels. The system's performance is likely to degrade in the presence of additive noise modulated at frequencies close to voice pitch. In preliminary tests, the formant-tracking subsystem proved to be reasonably robust over other values of the source spectrum slope (m) in addition to -9 dB/octave (6). However, it has shown sensitivity to the slope offset parameter (k). Smaller values of this parameter led to a lack of contrast between modulation strengths across filter channel outputs, hence resulted in the loss of minima corresponding to formant frequencies. In addition, increasing the value of this parameter would result in an increase of F1 over-estimation errors in vowels with low F1 frequencies (e.g. /ee/) for reasons explained previously.

The purpose of the formant enhancement stage is to selectively boost single harmonics closest to $F1_{est}$ and $F2_{est}$. The bandwidth of the FIR filters used was set to 50 Hz, however the most suitable value for this parameter will be known through subjective evaluation experiments. For the same gain, a larger bandwidth is likely to be perceived as louder and less tone-like. However, increasing the bandwidth beyond values close to F0 result in audible fluctuations near formant frequencies due to the increased interference from adjacent harmonics.

During preliminary testing, a subject with high frequency hearing loss was allowed to listen to a few sentences processed by the vowel-enhancement system in order to adjust the volume to a comfortable level. The subject was then presented with sentences at values of g_1 and g_2 (see section B. in Methods) spanning 0 dB to 21 dB and the range of acceptable gains was determined. For this particular subject, the preferred range was between 6 dB and 15 dB. The subject was then presented a wider range of sentences processed using these gain parameters. The subject described the processed sounds as being noticeably different compared to reference sentences (processed with zero gains) but acceptable and "sharper", for gains of 6 dB and 9 dB. The subject also described some sounds as being shrill and reported the perception of twin-voices. This perception of a "chorus-like" effect is likely due to phase distortion at the skirts of the narrowband bandpass filters used and better filter design or an appropriately wide bandwidth may lead to a more natural perceived quality.

During testing with sentences, low-frequency artifacts related to the frame length were audible, likely due to discontinuities at frame boundaries resulting from the formant enhancement stage. Further work is warranted in order to alleviate these problems in any frame-based implementation of the vowel-enhancement system. A sample-based system was created for generating test sentences for subjective testing. In this system, pitch and formant-tracking were done on a full sentence input using their frame-based implementations. However, smoothly varying values of $F0_{est}$, $F1_{est}$ and $F2_{est}$ were obtained by interpolating between each frame's estimates. Using these parameters, the coefficients of the FIR filters in the formant enhancement stage were updated for each sample. This successfully removed the low-frequency noise due to frame discontinuities; however occasional noise artifacts that can be described as descending or ascending pure-tones were audible. This artifact was found to arise when any slowly changing $F0_{est}$ is nearly equidistant from two adjacent harmonics and crosses over to the other half, making it closer to the harmonic adjacent to the one being currently boosted. Due to this crossover, the harmonic adjacent to the previously-dominant harmonic is now boosted, resulting in an abrupt and perceivable artifact. This noise was especially noticeable due to the "musical-noise"–like nature of this artifact. Two steps were taken to tackle this problem. First, calculation of the harmonic closest to $F1_{est}$ and $F2_{est}$ was modified to include history of the last harmonic calculated. Using this information, abrupt jumps between adjacent harmonics detected and the harmonic calculation was modified to select the harmonic closest to the harmonic calculated previously. Although this would lead to higher deviation from the true harmonic being boosted, it would still serve the purpose, because the gains applied to harmonics are likely to be high enough to make the harmonic dominate the AN fiber bandpass filter. This step is successful in eliminating most of the jumps between adjacent harmonics. The second step was to identify the remaining jumps between harmonics and activate a transitory phase in which the gain of the filter centered at the new harmonic was slowly increased to its full value while the gain of the filter centered at the old harmonic was slowly reduced. Exponential ramping was used to monotonically modify the gains. An appropriate length of the transitory phase was found to be 300 samples. Although this sample-based system is not suited for real-time use, it is likely to prove useful for testing the system on real-world sentences and serve as a useful proof-of-concept.

IV. Conclusion

A new signal-processing strategy based on recent neurophysiological observations was developed with the aim of improving vowel discrimination in listeners with hearing loss. The observations also guided the design of a novel formant-tracking strategy which showed reasonable ability to generalize over multiple speakers and vowels. Objective evaluation of the formant-tracking strategy was carried out

and described. Future areas of improvements were identified.

REFERENCES

[1] R. A. Cole, Y. Yan, B. Mak, M. Fanty, and T. Bailey, "The contribution of consonants versus vowels to word recognition in fluent speech," in *Proc. IEEE ICASSP*, Atlanta, GA, May. 1996. pp. 853-856.

[2] D. Kewley-Port, T. Z. Burkle, and J. H. Lee, "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 122, no. 4, pp. 2365-75, Oct. 2007.

[3] M. J. Owren, and G. C. Cardillo, "The relative roles of vowels and consonants in discriminating talker identity versus word meaning," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1727-39, Mar. 2006.

[4] G. Parikh, and P. C. Loizou, "The influence of noise on vowel and consonant cues," *J. Acoust. Soc. Am.*, vol. 118, no. 6, pp. 3874-3888, Dec. 2005.

[5] G. Fant, *Acoustic Theory of Speech Production*, The Hague, Netherlands: Mouton, 1960.

[6] D. B. Fry, A. S. Abramson, P. D. Eimas, and A. M. Liberman, "The Identification and Discrimination of Synthetic Vowels," *Lang. Speech*, vol. 5, no. 4, pp. 171-189, Oct. - Dec. 1962.

[7] H. L. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*: Cambridge University Press, 2009.

[8] M. Ito, J. Tsuchida, and M. Yano, "On the effectiveness of whole spectral shape for vowel perception," *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1141-1149, Aug. 2001.

[9] S. A. Zahorian, and A. J. Jagharghi, "Spectral-shape features versus formants as acoustic correlates for vowels," *J. Acoust. Soc. Am.*, vol. 94, no. 4, pp. 1966-1982. 1993.

[10] J. D. Miller, "Auditory-perceptual interpretation of the vowel," *J. Acoust. Soc. Am.*, vol. 85, no. 5, pp. 2114-2134, May. 1989.

[11] R. K. Potter, and J. C. Steinberg, "Toward the Specification of Speech," *J. Acoust. Soc. Am.*, vol. 22, no. 6, pp. 807-820. 1950.

[12] B. Mohr, and W.-Y. Wang, "Perceptual distance and the specification of phonological features," *Phonetica*, vol. 18, no. 1, pp. 31-45. 1968.

[13] L. C. W. Pols, L. J. T. v. d. Kamp, and R. Plomp, "Perceptual and Physical Space of Vowel Sounds," *J. Acoust. Soc. Am.*, vol. 46, no. 2B, pp. 458-467, Aug. 1969.

[14] R. N. Shepard, "Psychological representation of speech sounds," *Human communication: A unified view. New York: McGraw-Hill*, pp. 67-113. 1972.

[15] L. H. Carney, and J. M. McDonough, "Neural representations of vowels and vowel-like sounds," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 3309-3309, Apr. 2012.

[16] E. D. Young, and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Am.*, vol. 66, no. 5, pp. 1381-1403, Nov. 1979.

[17] B. Delgutte, and N. Y. S. Kiang, "Speech coding in the auditory nerve: I. Vowel-like sounds," *J. Acoust. Soc. Am.*, vol. 75, no. 3, pp. 866-878, Mar. 1984.

[18] G. A. Studebaker, R. L. Sherbecoe, D. M. McDaniel, and C. A. Gwaltney, "Monosyllabic word recognition at higher-than-normal speech and noise levels," *J. Acoust. Soc. Am.*, vol. 105, no. 4, pp. 2431-2444, Apr. 1999.

[19] G. Langner, and C. E. Schreiner, "Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms," *J. Neurophysiol.*, vol. 60, no. 6, pp. 1799-1822, Dec. 1988.

[20] B. S. Krishna, and M. N. Semple, "Auditory Temporal Processing: Responses to Sinusoidally Amplitude-Modulated Tones in the Inferior Colliculus," *J. Neurophysiol.*, vol. 84, no. 1, pp. 255-273, Jul. 2000.

[21] P. C. Nelson, and L. H. Carney, "Neural Rate and Timing Cues for Detection and Discrimination of Amplitude-Modulated Tones in the Awake Rabbit Inferior Colliculus," *J. Neurophysiol.*, vol. 97, no. 1, pp. 522-539, Jan. 2007.

[22] G. Langner, "Periodicity coding in the auditory system," *Hearing Res.*, vol. 60, no. 2, pp. 115-142, Jul. 1992.

[23] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust.*, vol. 24, no. 5, pp. 399-418. 1976.

[24] J.-S. R. Jang, "Speech and Audio Processing (SAP) Toolbox", Available: http://mirlab.org/jang/matlab/toolbox/sap.

[25] S. Nooteboom, *The prosody of speech: melody and rhythm*, W. J. Jardcastle and J. Laver ed., Cambridge, MA: Blackwell, 1997.

[26] P. McLeod, and G. Wyvill, "A smarter way to find pitch," in *Proc. ICMC*. 2005.

[27] C. A. Shera, J. J. Guinan, and A. J. Oxenham, "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 5, pp. 3318-3323, Mar. 2002.

[28] M. F. Schwartz, "The acoustics of normal and nasal vowel production," *Cleft Palate J.*, vol. 5, pp. 125-40, Apr. 1968.

[29] M. S. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney, "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.*, vol. 126, no. 5, pp. 2390. 2009.

[30] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, vol. 97, no. 5 Pt 1, pp. 3099, May. 1995.

[31] R. L. Miller, B. M. Calhoun, and E. D. Young, "Contrast enhancement improves the representation of/[bold ε]/-like vowels in the hearing-impaired auditory nerve," *J. Acoust. Soc. Am.*, vol. 106, pp. 2693. 1999.

[32] L. Deng, L. J. Lee, H. Attias, and A. Acero, "Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model," *IEEE Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 13-23. 2007.

[33] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust.*, vol. 22, no. 2, pp. 135-141. 1974.

[34] B. S. Atal, and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.*, vol. 50, no. 2B, pp. 637-655, Aug. 1971.

[35] J. L. Flanagan, "Automatic Extraction of Formant Frequencies from Continuous Speech," *J. Acoust. Soc. Am.*, vol. 28, no. 1, pp. 110-118, Jan. 1956.

[36] A. Potamianos, and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc. Am.*, vol. 99, pp. 3795. 1996.

[37] R. W. Schafer, and L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," *J. Acoust. Soc. Am.*, vol. 47, no. 2B, pp. 634-648, Feb. 1970.

[38] B. Delgutte, "Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds," *J. Acoust. Soc. Am.*, vol. 75, no. 3, pp. 879-886, March. 1984.

[39] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh, "Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech," *IEEE Speech Audio Process.*, vol. 13, no. 5, pp. 776-786. 2005.

[40] J. A. Haigh, and J. S. Mason, "Robust voice activity detection using cepstral features," in *Proc. IEEE TENCON*, China, Oct. 1993. pp. 321-324.

[41] G. Hu, and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 8, pp. 2067-2079. 2010.

Akshay Rao received the Bachelor of Technology degree in information technology from Indraprastha University, New Delhi, India, in 2008 and the Master of Science degree in electrical and computer engineering from the University of Rochester, Rochester, NY, in 2013.

From 2008 to 2011, he was a Software Engineer at Tally Solutions Private Limited in Bangalore, Karnataka, India. He worked as a Hearing Science Research Intern at Starkey Hearing Research Center, Berkeley, CA in 2012. He has been a Software

Engineer (DSP) at Bose Corporation, Framingham, MA since 2013.

Mr. Rao was awarded the J. N. Tata Endowment and the Bharat Petroleum Corporation Limited scholarships for graduate studies outside India in 2011 and the Jamsetji Tata Scholarship in 2012.

Laurel H. Carney (M'13) received the S.B. degree in electrical engineering from Massachusetts Institute of Technology, Cambridge, MA, USA, in 1983 and the M.S. and PhD in electrical engineering from the University of Wisconsin-Madison, Madison, WI, in 1985 and 1989.

She was an Assistant/Associate Professor of Biomedical Engineering at Boston University from 1991-2001, and Professor of Biomedical Engineering and Neuroscience at Syracuse University from 2001-2007. Since 2007, she has been a Professor of Biomedical Engineering and of Neurobiology & Anatomy at the University of Rochester, Rochester, NY, USA.

Dr. Carney is a fellow of the Acoustical Society of America, and a member of the Association for Research in Otolaryngology, Biomedical Engineering Society, American Society for Engineering Education, American Auditory Society, and the Society for Neuroscience. She is a fellow of the American Institute for Medical and Biological Engineering.