Speaker normalization using cortical strip maps: A neural model for steady state vowel categorization

Heather Ames and Stephen Grossberg[a)]
Department of Cognitive and Neural Systems
Center for Adaptive Systems
and
Center of Excellence for Learning In Education, Science, and Technology
Boston University
677 Beacon Street
Boston, MA 02215

Journal of the Acoustical Society of America, in press

Technical Report CNS/CAS-TR-08-005
Submitted: November 24, 2007
Accepted: September 12, 2008

Running Title: Cortical speaker normalization for vowel categorization

[a)] Electronic mail for corresponding author: steve@bu.edu

Abstract:

Auditory signals of speech are speaker-dependent, but representations of language meaning are speaker-independent. The transformation from speaker-dependent to speaker-independent language representations enables speech to be learned and understood from different speakers. A neural model is presented that performs speaker normalization to generate a pitch-independent representation of speech sounds, while also preserving information about speaker identity. This speaker-invariant representation is categorized into unitized speech items, which input to sequential working memories whose distributed patterns can be categorized, or chunked, into syllable and word representations. The proposed model fits into an emerging model of auditory streaming and speech categorization. The auditory streaming and speaker normalization parts of the model both use multiple strip representations and asymmetric competitive circuits, thereby suggesting that these two circuits arose from similar neural designs. The normalized speech items are rapidly categorized and stably remembered by Adaptive Resonance Theory circuits. Simulations use synthesized steady-state vowels from the Peterson and Barney [J. Acoust. Soc. Am. 24, 175-184 (1952)] vowel database and achieve accuracy rates similar to those achieved by human listeners. These results are compared to behavioral data and other speaker normalization models.

PACS numbers: 43.71.An, 43.71.Es, 43.72.Bs

I. INTRODUCTION: SPEECH LEARNING, NORMALIZATION, AND IMITATION

Fundamental variations in speech exist both between speakers and within the speech of a single speaker. Intra-speaker variability is mainly concerned with the different pronunciations of the same phoneme by a single speaker. These variances can result from differences in phonemic context, including coarticulation effects, accent, and the emotions or stress level of the speaker. Inter-speaker variability concerns the variation of speech across speakers and these variations generally have a much larger effect on perception (Nearey, 1989). Despite this variability, a listener is able to identify and understand speech spoken by different speakers on the first encounter with a speaker and on nearly the first utterance. It seems that, not only does the listener's brain learn to store a speaker-invariant representation of speech, but that somehow the speech encountered is transformed, or *speaker normalized*, into a speaker-invariant representation for the purpose of understanding.

The process of speaker normalization enables a baby to begin to imitate sounds from adult speakers, notably parents whose spoken frequencies differ significantly from those that the baby can babble. A circular reaction from endogenously babbled to heard sounds enables a baby to learn a map between the auditory representations of its own heard babbled sounds to the motor commands that caused them (Piaget, 1963; Grossberg, 1978; Cohen *et al.,* 1988; Bullock *et al.,* 1993; Guenther, 1995; Guenther *et al.,* 2006). Speaker normalization enables sounds from adult caretakers to be filtered by this learned map and to thereby enable the baby to begin to imitate and refine heard sounds in its own language productions. Learning in such an imitative map needs to remain active for many years in order to enable an individual's changing voice through puberty and adulthood to continue to activate and update this map.

Speaker normalization also enables language meanings that were learned from one teacher's voice to be readily understood when uttered by another speaker. More generally, speaker normalization helps the brain to overcome a combinatorial explosion that would otherwise occur if the brain needed to store every instance of every speaker utterance in order to understand language meaning.

A similar problem of combinatorial explosion is overcome by the visual cortex as it learns to recognize visually perceived objects in the world. In vision, object category representations are learned that are relatively insensitive to object size, location, and orientation on the retina (Bradski and Grossberg, 1995; Ito *et al.,* 1995). Such invariance is built up across several processing stages, with invariance only appearing in the inferotemporal cortex and beyond. Likewise, speaker normalization is just one stage in the development of rate- and speaker-independent representations of language meaning. For example, Boardman *et al.* (1999) and Grossberg *et al.* (1997) have modeled how rate-invariance may develop across several processing stages.

Although speaker normalization and rate-invariance are important for learning to speak and understand language, humans and other animals are also exquisitely sensitive to the voice quality and prosody of individual speakers. A similar dichotomy occurs during visual learning and recognition, where positionally-invariant object recognition categories coexist with cortical representations that enable manipulation of objects in space. In both audition and vision, this is accomplished through interactions across What and Where cortical processing streams (Ungerleider and Mishkin, 1982; Goodale and Milner, 1992; Hickok and Poeppel, 2007; Fazl, Grossberg, and Mingolla, 2008). Such interactions have been predicted to compute computationally *complementary* properties (Grossberg, 2000): The properties needed to compute one such property prevent a complementary property from being computed in the same cortical processing stream, and conversely. Interactions between cortical processing streams that compute such complementary properties enable them to overcome their complementary deficiencies and thus to generate adaptive and creative behaviors. Thus, the present article's focus on speaker normalization is in no way contradicted by the fact that the brain can process many individual features of speaker identity and prosody. In fact, one property of the present speaker normalization model is that, while it generates a speaker-independent representation, it also marks the speaker-dependent frequencies of the selected speaker from which speaker-dependent properties can be computed in a complementary processing stream.

Another important issue concerns the specificity or generality of speaker-independent language categories. In certain environments, it is essential to distinguish fine differences between speaker utterances that can change language meaning across speakers. In other environments, considerable variability across utterances can lead to a similar meaning. Our model clarifies how such variability can naturally emerge during incremental learning of individual language exemplars. Thus, the fact that humans and animals can sometimes distinguish individual speech exemplars does not imply that language learning is exemplar learning, with all the problems of combinatorial explosion and biologically-implausible exhaustive search that such a model can create. Recognition categories can, instead, be learned in a way that tracks the task demands and statistics of each unique environment and that seems to conjointly maximize category generality and minimize predictive error. Neurophysiological data in the visual cortex support this viewpoint (Spitzer *et al.,* 1988; Zoccolan *et al.,* 2007) in a manner predicted by neural models (Carpenter and Grossberg, 1987, 1991; Carpenter, 1997; Grossberg, 1999). Learned category prototypes in such models can represent either individual exemplars or abstract knowledge, as each learning situation uniquely demands. Vowel category simulations in the present article illustrate this property.

Speaker normalization is also an important technique used in engineering for building automatic speech recognition (ASR) systems. Vowel classification rates in ASR systems can be improved if features are speaker-normalized before classification (Nearey, 1989). The Formant Ratio Theory is a foundation for many speaker normalization techniques. The Formant Ratio Theory states that vowel quality depends on the log frequency intervals between formants (defined as ratios), and that shifting activations along a log frequency axis will generate the invariant representation (Lloyd, 1890a, 1890b, 1891, 1892; Peterson, 1961; Bladon *et al.*, 1984; Sussman, 1986; Sydral and Gopal, 1986; Miller, 1989; Sussman *et al.*, 1997). Formant ratios can be calculated by averaging across formant values for many utterances of a single speaker.

Despite its heuristic appeal, in its classical formulation, Formant Ratio Theory faces two types of problems. First, no mechanism has been proposed to explain how the human auditory system could perform these calculations. How does the brain align cell activities corresponding to the formant frequencies for each utterance? Second, this method requires information contained in many speech samples of a single speaker and thus is not able to account for our ability to understand a speaker in the first utterance that we encounter. It is not biologically feasible for the brain to perform computations across all speech samples it encounters from a speaker, store this information, and then use it to normalize each new utterance encountered for each speaker.

The inability of ASR systems to understand speech in real situations and environments may be due to their lack of adherence to biological auditory principles. As Dusan and Rabiner (2005) pointed out, perhaps it is now time to take a closer look at how the brain performs speech recognition and apply these insights to design novel ASR systems. The modeling work presented in this paper proposes a new method for speaker normalization that makes use of the functional architecture of the brain and builds upon previous modeling work that explains a large amount of data in acoustics, speech perception, and language.

The well-documented existence of tonotopic organization in the auditory cortex, which gives rise to *strips* of frequency selective cells, serves as the functional architecture within which the speaker normalization transformation is proposed to occur; see Section II. Frequency-selective strips provide more representational space within which finer computations can occur. They share some properties with the *hypercolumn* organization that is ubiquitous in the visual cortex (Hubel and Wiesel, 1962). In vision, hypercolumns occur in cortical maps that represent multiple features of visual objects in physical space. In audition, they occur in cortical maps that represent multiple features of acoustic objects in frequency space.

Such *strip maps* have earlier been shown capable of explaining key data about auditory streaming, or the separation of acoustic sources (Grossberg *et al.*, 2004). Speaker normalization and streaming circuits may have arisen from similar underlying neural designs. In particular, the streaming model circuit and the speaker normalization circuit described herein use both strip maps and asymmetric competition across frequency-selective channels to realize a kind of "exclusive allocation." During auditory streaming, exclusive allocation enables spectral information to be allocated to a specific source or stream (Bregman, 1990). Here we predict that the exclusive allocation properties that are familiar in streaming research also play a role in generating a speaker-independent representation of speech. Strip maps have also been used to explain other cortical processes, such as how place-value number systems may be learned (Grossberg and Repin, 2003). Strip maps may thus be a cortical design that has been specialized during brain evolution to accomplish multiple tasks.

As explained below, a simple transformation from speaker-dependent to speaker-independent speech information can be performed within strip maps in a way that is consistent with neurobiological data. The speaker-independent representations are then categorized via a process of fast incremental learning. Results from synthesized steady-state vowel categorization simulations are presented to validate the performance of the speaker normalization model. The results have been briefly reported in Ames and Grossberg (2006, 2007).

II. TONOTOPIC ORGANIZATION AND MULTIPLE STRIP MAPS

The auditory system contains spatially organized maps of frequency selective cells called tonotopic maps. The frequency representations are arranged logarithmically. Tonotopy is preserved in the auditory system from the level of the cochlea to the auditory cortex of humans and other mammals (Tunturi, 1952; Merzenich and Brugge, 1973; Imig *et al.*, 1977; Reale and Imig, 1980; Romani *et al.*, 1982; Seldon, 1985; Luethke *et al.*, 1988; Pantev *et al.*, 1988; Morel and Kaas, 1992; Morel *et al.*, 1993; Cansino *et al.*, 1994; Heil *et al.*, 1994; Rauschecker *et al.*, 1995; Bilecen *et al.*, 1998; Wessinger *et al.*, 1998; Lockwood *et al.*, 1999; Talavage *et al.*, 2000, 2004; Rauschecker and Tian, 2004). In the auditory cortex, these tonotopic maps consist of iso-frequency contours which can be defined as *strips* of cortical cells that respond to a specific frequency, or best frequency.

In addition to spectral information that is explicitly in acoustic inputs, missing fundamental frequencies (F0) of harmonic sounds activate the tonotopic maps of the primary auditory cortex of mammals. Single-unit extracellular recordings in marmosets have shown that complex tones with missing fundamentals activate tonotopic areas corresponding to the missing fundamental (Bendor and Wang, 2005). These maps were found in the low frequency-selective areas on the border of core areas AI and R and the lateral belt areas AL and ML, but did not extend into the entire tonotopic representation of any of these areas. Fishman *et al.* (1998) found an implicit representation of the missing fundamental in AI based on population neuronal responses in awake macaque monkeys. Missing fundamental activations have also been seen in auditory cortical areas of gerbils (Schulze *et al.*, 2002) and cats (Whitfield, 1980; Qin *et al.*, 2005).

In humans, fMRI has been used to show that the lateral Heschl's gyrus is sensitive to the F0 differences of iterated rippled noise (IRN) when subjects listened to noise with temporally varying patterns (Patterson *et al.*, 2002). Penagos *et al.* (2004) confirmed the existence of this F0-selective region by using fMRI to show that missing fundamental complex tones containing only low frequency harmonics causes a stronger activation in this region than if the tones contained only high frequency harmonics. This difference is attributed to the unresolvability of the high frequencies for listeners. Langner *et al.* (1997) found that a topographically ordered F0 map in human auditory cortex (where F0 was described as the periodicity of the complex sound) may be found orthogonally to the topographically ordered spectral map.

These data confirm that cells in auditory cortex respond selectively to frequencies and F0 in a spatially organized manner, but the exact placement of an F0-sensitive map with respect to a spectral map is unclear. For the purpose of our speaker normalization model, it is assumed that the F0-sensitive map may lie near or within the spectrally activated maps. Simulations were performed to manipulate the amount energy at the F0 filters in order to test the role of F0 in the speaker normalization transformation. The advantages and disadvantages of using F0 for speaker normalization are discussed below.

Our speaker normalization model builds upon the fact that *multiple* tonotopic maps of frequency-selective strips are found in the auditory cortex of both humans and other mammals (Merzenich and Brugge, 1973; Imig *et al.*, 1977; Morel and Kaas, 1992; Morel *et al.*, 1993; Hacket *et al.*, 1998; Kaas and Hackett, 1998; 2000; Formisano *et al.*, 2003; Rauschecker and Tian, 2004; Petkov *et al.*, 2006). Interactions between such maps are core design features of our speaker normalization model. Map boundaries are defined by frequency reversals such that the low frequency endpoint of one map is adjacent to the low frequency endpoint of the next map. The same occurs for the high frequency endpoints. Talavage *et al.* (2004) used fMRI and frequency-swept stimuli to identify six tonotopic mappings in the superior temporal plane, suggesting that there are at least five areas in the human auditory cortex that exhibit at least six tonotopic organizations. However, the number of maps that exist in the human brain is still uncertain and is difficult to determine with the resolution available in imaging technologies.

The speaker normalization model presented in this paper assumes that at least two of these tonotopic strip maps have an orthogonal, or at least non-parallel, spatial arrangement. The overlapping interactions between these two spectral maps allow the spectral information from different speakers to be aligned along a *diagonal map*. This diagonal map arrangement underlies the computations needed to shift the speaker-dependent speech information into a speaker-independent representation.

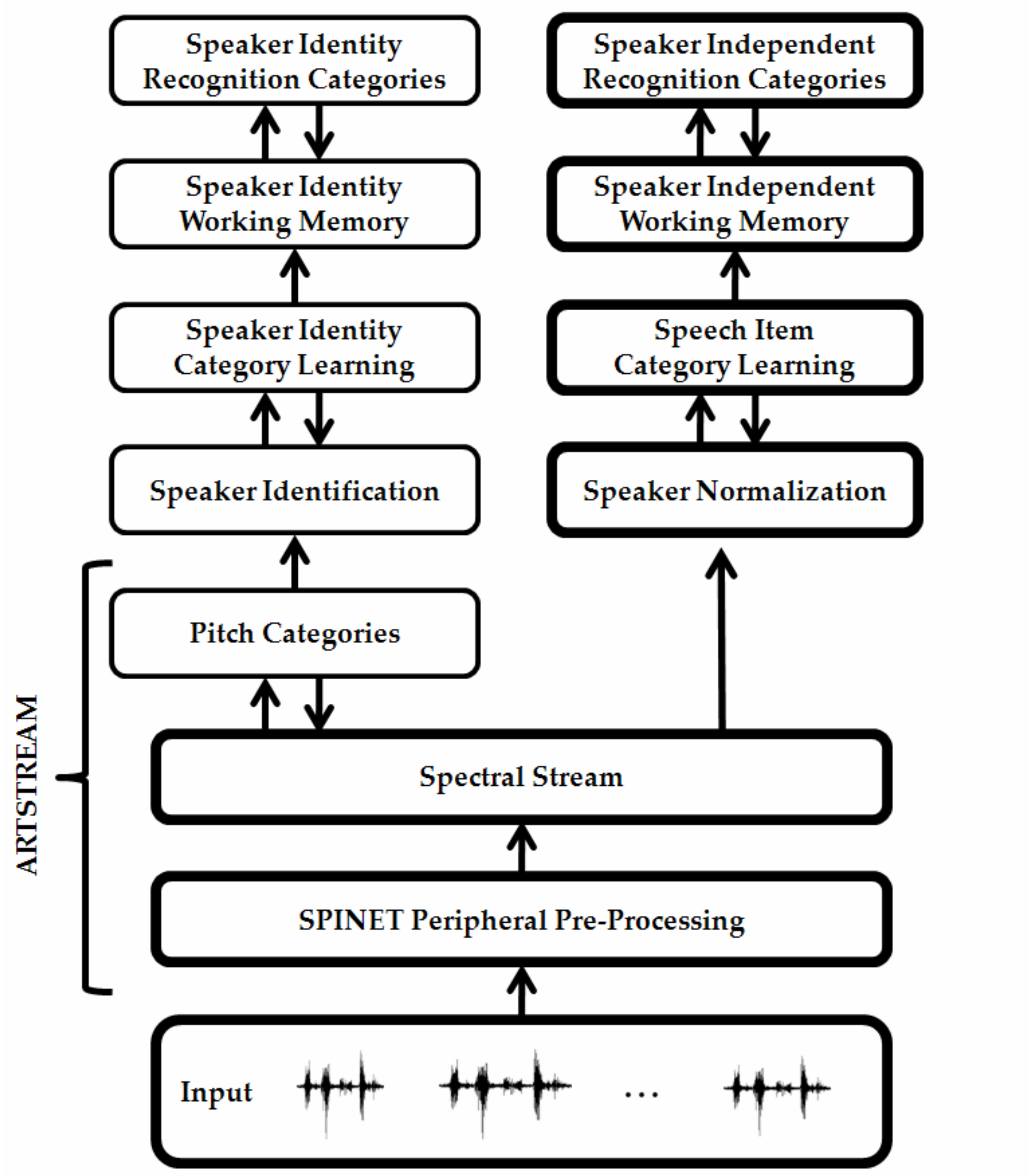
Speaker Identity Recognition Categories
Speaker Independent Recognition Categories
Speaker Identity Working Memory
Speaker Independent Working Memory
Speaker Identity Category Learning
Speech Item Category Learning
Speaker Identification
Speaker Normalization
Pitch Categories
ARTSTREAM
Spectral Stream
SPINET Peripheral Pre-Processing
Input
...

Figure 1: *Box diagram of the ARTSPEECH perception system.* The boldface boxes contain components discussed or simulated in this paper.

III. AN EMERGING AUDITION, SPEECH, AND LANGUAGE MODEL

The model presented in this paper called the Neural Normalization Network, or NormNet for short. NormNet is part of an architecture for speech perception and recognition that is being developed by Grossberg and colleagues (Cohen *et al.*, 1995; Cohen and Grossberg, 1997; Grossberg *et al.*, 1997; Boardman *et al.*, 1999; Grossberg and Myers, 2000; Grossberg, 2003b; Grossberg *et al.*, 2004); see Figure 1. At the periphery of this architecture, a Spatial Pitch NETwork (SPINET) processes acoustic information and converts the temporally-occurring auditory signals into spatial representations of pitch (Cohen *et al.*, 1995; see Figure 2). Harmonically-related spectral components (see Stages 6 and 7 in Figure 2) can activate a given pitch category through an adaptive filter. The selection of harmonics is due to learning that is driven by the natural grouping of frequencies in early auditory processing. SPINET hereby creates both spatial representations of pitch and harmonically related spatial activations. This mapping is a crucial feature for the proposed speaker normalization technique.

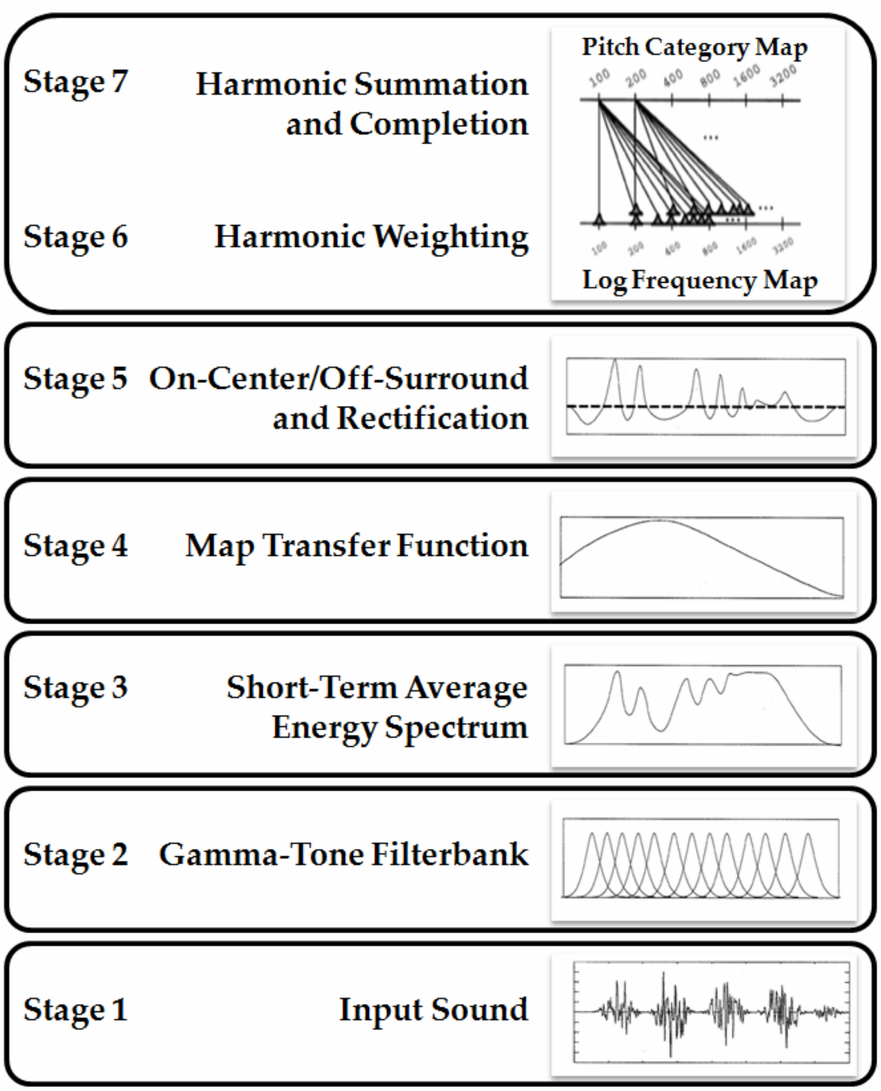

Figure 2: *SPINET Model*. The processing stages transform a sound stream into activations of spatially distributed pitch nodes. [Reprinted with permission from Cohen, Grossberg, and Wyse 1995]

SPINET provides a natural front end for a more comprehensive model of pitch-based auditory streaming that is called the ARTSTREAM model (Grossberg *et al.*, 2004; see Figure 3). Both the spectral and pitch representations in SPINET are defined by strips of frequency and pitch. The frequency strips in the spectral maps are selective for a particular frequency and are ordered on a log frequency axis. These frequency selective strips are a key organizational structure in ARTSTREAM that allows the model to parse acoustical information into distinct auditory streams that intersect the strips at an orthogonal, or at least non-parallel, angle.

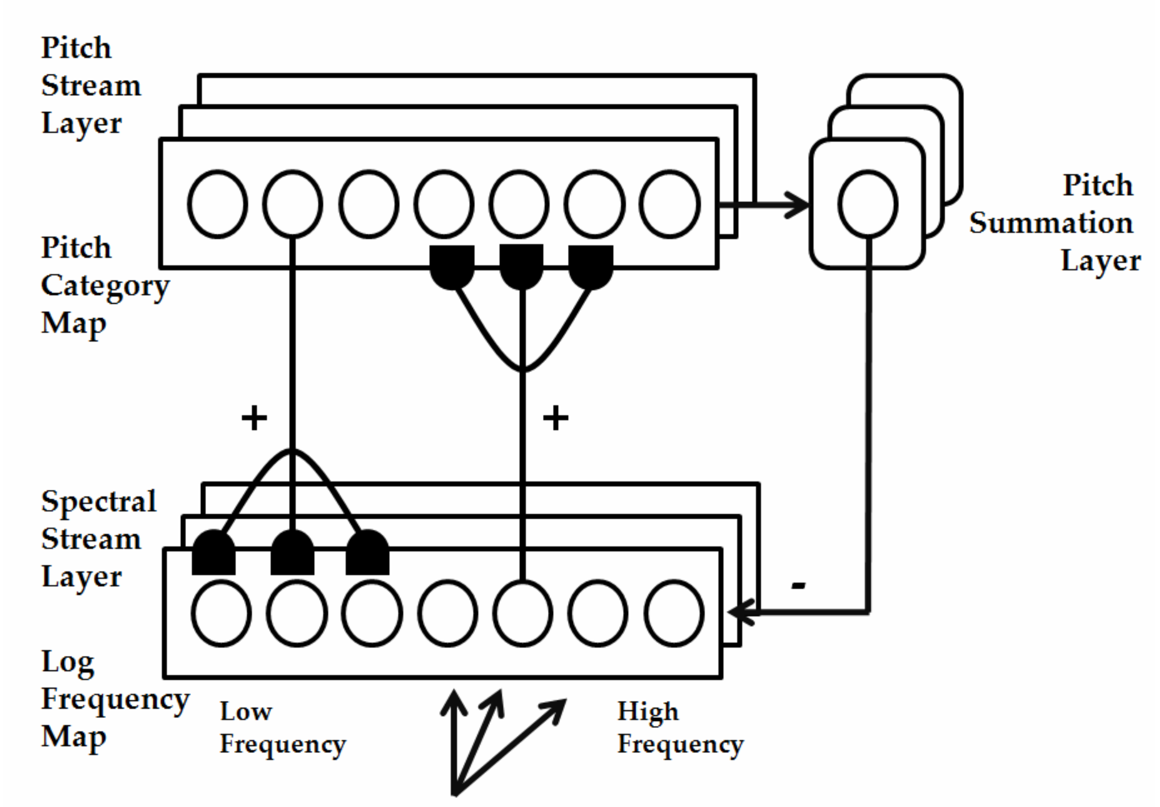

Figure 3: *ARTSTREAM Model.* The spectral and pitch layers of the SPINET model (layers 6 and 7) are elaborated in the ARTSTREAM Model into multiple representations, or strips of cells, and top-down ART matching also occurs. Bottom-up signals group harmonically-related spectral components into activations of pitch categories. Inhibition within each pitch stream enables only one pitch category to be active at any time in a given stream. Asymmetric inhibition across streams in the pitch stream layer is biased so that the winning pitch cannot be represented in another stream. The winning pitch category feeds back excitation to its harmonics in the corresponding spectral stream. This stream also receives nonspecific top-down inhibition from the pitch layer. ART matching is hereby realized. It suppresses those spectral components that are not harmonically related to the active pitch. Inhibition across spectral streams then prevents the resonating frequency from being represented in other streams as well. [Reprinted with permission from Grossberg 2003]

ARTSTREAM derives its name from Adaptive Resonance Theory, or ART (Grossberg, 1976a, 1976b, 1978, 1980). ART principles and mechanisms have been used to explain and predict data about visual and auditory perception, category learning and object recognition, cognitive information processing, cognitive and emotional interactions, and their underlying brain mechanisms (Carpenter and Grossberg, 1991; Grossberg, 1994, 1999, 2003a; Grossberg and Merrill, 1996; Chey *et al.*, 1997; Grossberg, Boardman, and Cohen, 1997; Grunewald and Grossberg, 1998; Grossberg and Williamson, 1999; Vitevitch and Luce, 1999; Page, 2000; Grossberg and Myers, 2000; Bowers, 2002; Goldinger and Azuma, 2003; Hawkins, 2003; Fazl *et al.,* 2008; Grossberg and Versace, 2008). ART claims that resonant states between top-down expectations and bottom-up input drive stable learning of perceptual and cognitive representations, while preventing catastrophic forgetting of previously learned information.

In the domain of audition, speech perception, and language, models based on ART mechanisms have been used to explain, in addition to auditory streaming, word recognition and recall (Grossberg and Stone, 1986), manner distinctions in consonant perception (Boardman *et al.*, 1999), consonant integration and segregation in VC-CV syllables (Grossberg *et al.*, 1997), and interword integration and duration-dependent backward effects (Grossberg and Myers, 2000). These models mechanistically embody such design principles as storage in working memory of temporal order information derived from phonemic representations, automatic gain control to maintain rate invariance, and top-down matching of learned expectations with bottom-up patterns of information in order to focus attention on expected combinations of acoustic features and to stabilize fast auditory learning. See Grossberg (2003b) for a review.

ARTSTREAM includes a bottom-up adaptive filter, or "harmonic sieve," that groups together harmonics of an auditory source into learned pitch categories. In addition, a top-down filter encodes the expectations of the learned pitch categories. Each expectation consists of the harmonics of the learned pitch category, which competitively inhibit other frequencies. Both psychological and neurobiological data support the existence of such "biased competition" in the selection of attended data (Grossberg, 1980; Desimone, 1998; Grossberg, 1999; Kastner and Ungerleider, 2001; Grossberg, 2003a). An auditory stream forms when a bottom-up adaptive filter and its top-down expectation interact to generate a spectral-pitch resonant state. Through such resonant dynamics, ARTSTREAM is able to coherently select pitch-consistent frequencies, corresponding to both F0-related harmonics and formant frequencies, while suppressing other frequencies. This ARTSTREAM process along with asymmetric competition across streams realizes the property of "exclusive allocation" (Bregman, 1990). Although ARTSTREAM has discussed only the special case of how pitch categories may be learned from spectral information and thereby used to separate distinct acoustic sources, the same ART mechanisms can learn to categorize other speaker-specific properties that can be used for speaker identification.

The spectral information in the selected stream can be the input to the speaker normalization model, since the spectral-pitch resonances, or other speaker-specific resonances, isolate different speaker's sounds from one another; see Figure 1. Moreover, the spatially organized frequency-selective strips of ARTSTREAM provide the computational substrate that is needed to initiate speaker normalization. The key design principle of frequency-selective strip maps allows these models to seamlessly connect and interact. Isolated vowels are the inputs to NormNet in the current simulations, so the stream-separating mechanisms are neither needed nor simulated.

After speaker normalization is accomplished and the invariant vowels are categorized by an ART network, the speaker-independent vowel categories are in a form that can naturally input to the ARTWORD model of variable-rate speech categorization and word recognition (Grossberg and Myers, 2000).

IV. MODEL DESCRIPTION

Peripheral processing in the NormNet model is based on the SPINET model (Figure 2) of Cohen *et al.*, (1995) with a few modifications. The gammatone filterbank (see Stage 2 in Figure 2) consists of a cascade of fourth order gammatone filters (Holdsworth *et al.*, 1988; Patterson *et al.*, 1988; Cohen *et al.*, 1995):

$$GT(f) = [1 + j(f - f_i)/b(f_i)]^{-4} . \qquad (1)$$

The center frequencies (f_i) of the filters range from 10 to 8000 Hz and are equally spaced in ERB (equivalent rectangular bandwidth) units (Patterson and Rice, 1987; Patterson *et al.*, 1987;

Patterson *et al.*, 1988; Holdsworth *et al.*, 1988; Slaney, 1998; Slaney, 1993). The dynamic range corresponds to data measuring the dynamic range in human listeners (Hudspeth, 2000; Plack and Oxenham, 2005). The ERB of a filter at the center frequency (f_i) is a function of the filter center frequency (Glasberg and Moore, 1990):

$$ERB(f_i) = 24.7 + 0.108 * f_i, \tag{2}$$

and the bandwidth $b(f_i)$ of a filter is defined by:

$$b(f_i) = \frac{ERB(f_i)}{0.982} \ . \tag{3}$$

The output signal from the filterbank is then mapped onto a logarithmic scale, half-wave rectified, and low-pass filtered. This signal serves as the input to the speaker normalization model.

The speaker normalization transformation is proposed to occur in auditory cortex by using at least two intersecting tonotopic strip maps that are assumed, for simplicity, to align orthogonally. The names of these maps are the Anchor Log Frequency Map (*Anchor Map*) and the Stream Log Frequency Map (*Stream Map*); see Figure 4. Because both maps are composed of strips of frequency-selective units, the activations in these maps spread along the strips into an *inter-strip area* where strips from both maps are superimposed upon each other. Both the Anchor Map and Stream Map receive spectral information from the speech sound. In the full architecture, this spectral information is predicted to be the streamed output from a process like ARTSTREAM. In the current simplified model, SPINET preprocessing generates the model's spectral input pattern.

Asymmetric competition occurs in the Anchor Map to choose the cell with the lowest active frequency in the speech sound, which typically contains the largest amount of spectral energy (see Figure 4a). This cell is called the *anchor frequency coding cell.* As the anchor frequency coding cell wins the asymmetric competition, it inhibits any activations corresponding to higher frequencies in the Anchor Map. This form of "exclusive allocation" is predicted to be a key step in speaker normalization. The asymmetric competition is governed by the following on-center, off-surround shunting equation (Grossberg, 1973, 1980); see Figure 4a:

$$\frac{dx_{i0}}{dt} = -Ax_{i0} + (B - x_{i0})[I_{i0} + f(x_{i0})] - x_{i0} \sum_{i<k} f(x_{k0}), \tag{4}$$

where x_{i0} is the activity of the *ith* frequency-selective cell in the Anchor Map, and I_{i0} is the input to this cell in the Anchor Map. In equation (*4*), A = 0.1, B = 1, and $f(x) = x^2$. Since $f(x) = x^2$ is a faster-than-linearly increasing signal function, the activities of cells corresponding to the lowest frequency will increase as the activities of the other cells decrease, resulting in contrast enhancement and winner-take-all choice of the cell whose activity corresponds to the lowest active frequency (Grossberg, 1973).

The cell that codes the anchor frequency triggers coincidence detection along its strip in the inter-strip area where both the Anchor Map and the Stream Map activate their corresponding frequency-selective strips. The coincidence occurs in the strip corresponding to the Anchor Frequency of the Anchor Map (*ith* row) and all the active strips corresponding to spectral activations in the Stream Map (*jth* columns); see Figure 4b. The activity, x_{ij}, of the cell in the *ith* row and the *jth* column obey:

$$\frac{dx_{ij}}{dt} = -Ax_{ij} + g(x_{i0})I_j, \tag{5}$$

where x_{i0} is the activity of the i^{th} strip in the Anchor Map, I_j is the spectral representation of the speech sound at the j^{th} strip of in the Stream Map, the decay rate A= 0.1, and the Anchor Map sigmoid signal function

$$g(x_{i0}) = \frac{x_{i0}^b}{c^b + x_{i0}^b}, \tag{6}$$

where the choices b = 100 and c = 0.5 enable $g(x_{i0})$ to approximate 1 at the anchor frequency and 0 elsewhere. Due to coincidence detection $g(x_{i0})I_j$ in *(5)*, the Stream Map shifts into the Anchor Frequency Strip and becomes the *Anchored Stream*.

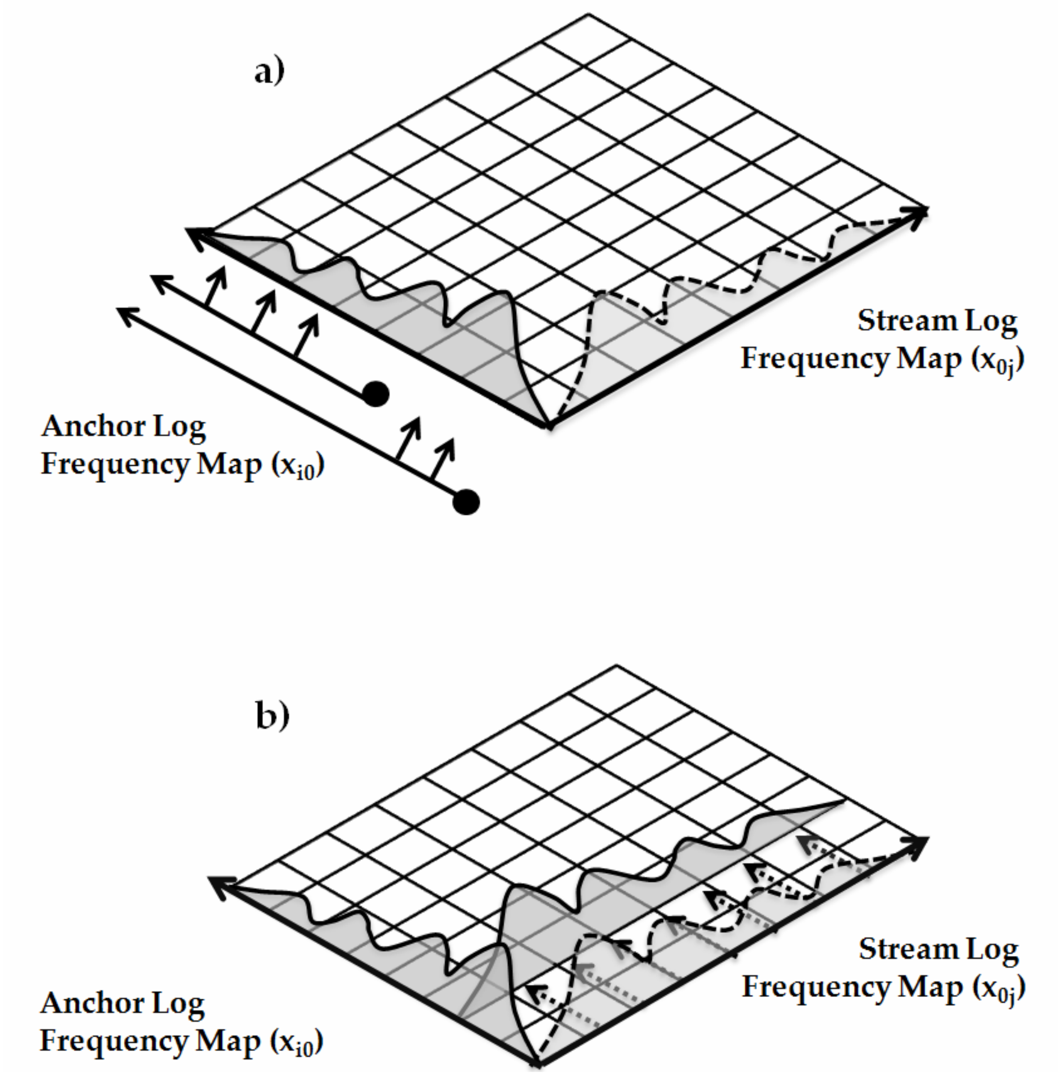

Figure 4: (a) *Anchor Map and Stream Map*. These Maps are organized orthogonally and superimpose on each other. Both maps receive spatially organized spectral information from the streamed sound. The activations spread along their corresponding strips into the inter-stream area. **(b)** *Coincidence detection*. The winning Anchor Frequency Coding Cell triggers a coincidence detection along its Anchor Frequency Strip. This coincidence detection moves the activations of the Stream Map into the Anchor Frequency Strip.

Because the Anchor Map and the Stream Map have connections which superimpose orthogonally, their coincidences can create diagonally connected strips; see Figure 5. These diagonal connections transform the Anchored Stream into a speaker-invariant representation, *S*. In particular, each cell in the *S* field sums inputs from all the cells along a diagonal created by the maps. The activity, s_m, of the m^{th} diagonal map cell is thus

$$s_m = \sum_{j=1}^{n} x_{j-m,j} \quad , \qquad (8)$$

where n is the number of filters in the gammatone filterbank and m is the cell number in the *S* field. The speaker-independent spectrum is then categorized into unitized item representations. These learned recognition categories are used for vowel identification.

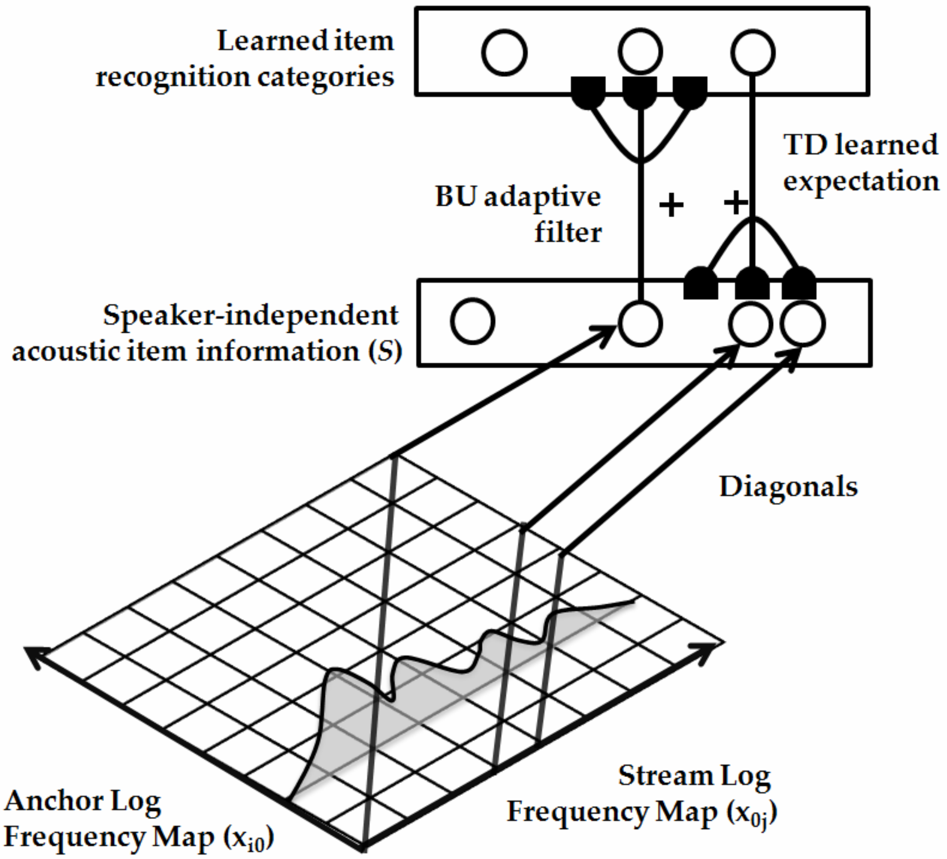

Figure 5: *Creation of speaker-independent working memory item information.* The diagonal strips are sampled to create the speaker-independent working memory item information which is then feed into an ART network which learns to categorize the item information.

In this paper, vowel categorization is carried out by a fuzzy ARTMAP network with default parameters (Carpenter *et al.*, 1992); see Figure 6. Fuzzy ARTMAP is a neural network that incorporates two fuzzy ART modules, ART$_a$ and ART$_b$, where ART$_a$ learns to map the speaker-independent vowel spectra to vowel categories. An intervening *map field*, F^{ab}, learns to associate the vowel categories to category names in ART$_b$. See the Appendix for the fuzzy ART equations.

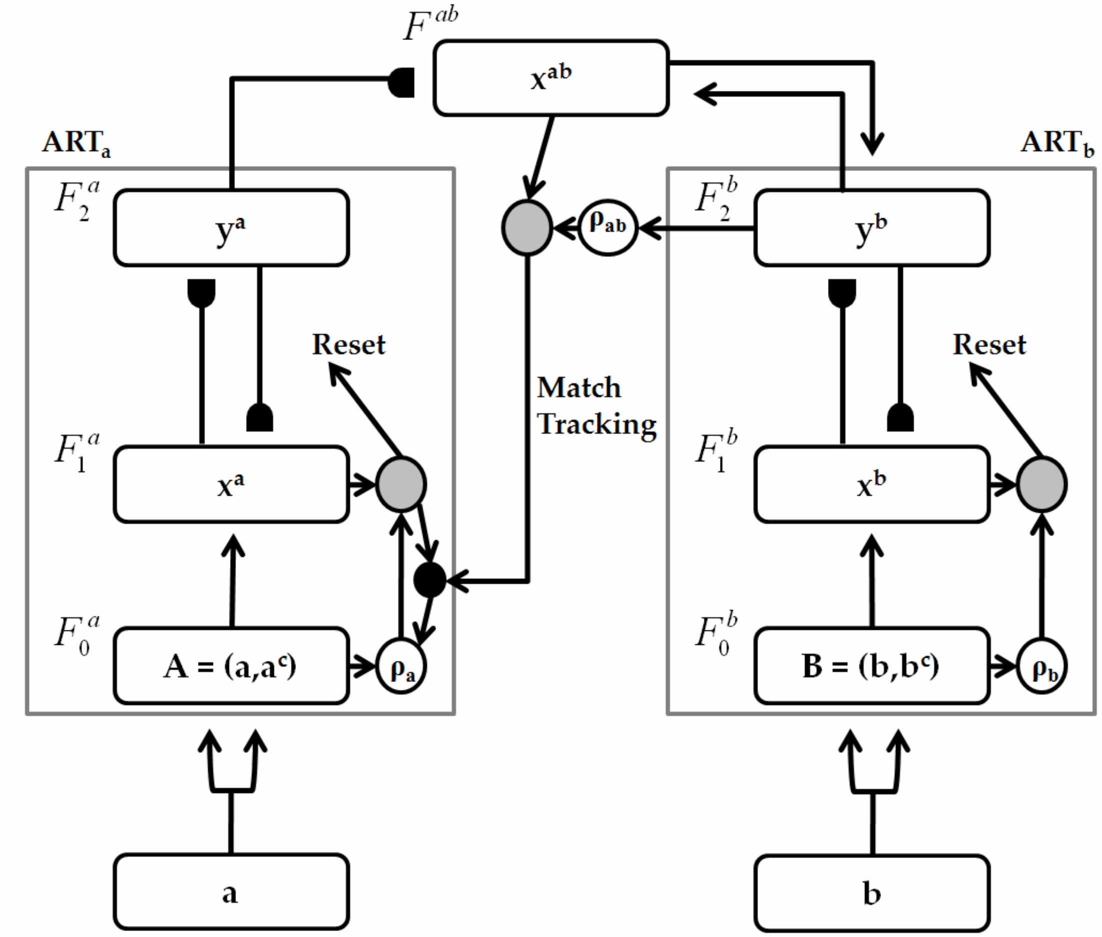

Figure 6: *Fuzzy ARTMAP*. The ART$_a$ complement coding preprocessor transforms the M_a vector **a** into the $2M_a$ vector **A** = (**a**,**a**^{c}) at the ART$_a$ field F_0^a. **A** is the input vector to the ART$_a$ field F_1^a. Similarly, the input to F_1^b is the $2M_b$ vector **B** = (**b**,**b**^{c}). When a prediction by ART$_a$ is disconfirmed at ART$_b$, inhibition of map field activation induces the match tracking process. Match tracking raises the ART$_a$ vigilance (ρ_a) to just above the F_1^a to F_0^a match ratio $|\mathbf{x}^a|/|\mathbf{A}|$. This triggers an ART$_a$ search which leads to activation of either an ART$_a$ category that correctly predicts **b** or to previously uncommitted ART$_a$ category node. [Reprinted with permission from Carpenter *et al.*, 1992]

When a vowel category is activated during category learning, it reads out a learned top-down expectation that is matched against the input vowel frequency spectrum. A *vigilance* parameter, ρ_a, determines whether the match is good enough. Learning occurs only if the match is good enough. Then a resonant state can develop that supports the learning process. A predictive failure in category naming at ART$_b$ increases ρ_a by the minimum amount needed to trigger a memory search. Such a memory search automatically leads to learning and/or selection of a new vowel category in ART$_a$ that can better match the vowel frequency spectrum. This process is called *match tracking* (Carpenter *et al.*, 1992). It enables learning of the most general vowel categories that can minimize predictive errors in ART$_b$. Match tracking is realized in the F^{ab} map field.

The speaker-independent spectral input vectors **A** to the F_1^a field of ART$_a$ are transformed into complement-coded vectors **A**=(**a**,**a**^{c}) before being further processed by ART$_a$. Complement coding means that both the activities **a** of the network's ON cells and the activities **a**^{c} = 1- **a** of its OFF cells form the input vector. The inputs **B** to F_1^b field of ART$_b$ are complemented-coded representations of vowel names: **B**=(**b**,**b**^{c}). The values of all components in these input vectors lie between 0 and 1. The activity in the F_1^a field activates a vowel category node, *J*, in the F_2^a field which, in turn, sends top-down signals to the F_1^a field, where matching between the bottom-up input and the top-down weight vector, $\mathbf{w}_J^a$ of the expectation occurs. If the match is good enough, as determined by the vigilance criterion:

$$\rho_a |A| < |A \wedge w_J| . \qquad (9)$$

then learning occurs. Inequality (9) describes the balance between excitation and inhibition at a novelty-sensitive orienting system such as the nonspecific thalamus or hippocampus (Carpenter and Grossberg, 1993; Grossberg and Versace, 2008). Term $\rho_a |A|$ describes the total amount of excitation that reaches the orienting system. Term $|A \wedge w_J|$ describes the total amount of inhibition that reaches the orienting system. In all, inequality (9) says that inhibition is greater than excitation, so that the orienting system does not fire. The matched patterns are thus allowed to resonate and learning is enabled. When excitation exceeds inhibition, the currently active category is reset and the network continues searching for a better category with which to encode the speaker-independent spectrum. The vigilance ρ_a value in (9) this term is a gain control term that determines the network's sensitivity to excitation from the bottom-up total input $|A|$. Increasing vigilance makes the network more sensitive to mismatches, and thus leads to finer categories. At very high vigilance, the network can learn individual exemplars. At low vigilance, it can learn abstract categories that enable many exemplars to be coded by the same recognition category.

During learning, the speaker-independent input pattern **A** is encoded by a vowel category in F_2^a , while the vowel name input pattern **B** is encoded by a name category *K* in F_2^b. In the present simulations, name category labels are directly input to F_2^b without loss of generality. The map field F^{ab} associates these categories unless *J* has previously learned to predict a different *K* category. If this occurs, then match tracking proceeds until an appropriate new ART$_a$ category is chosen and learned. During testing, the speaker-independent input signal **A** activates a name category in ART$_b$ through F^{ab}, which is the prediction of the system. Mathematical details about fuzzy ARTMAP are found in the Appendix.

V. METHODS

A. Stimuli

In order to evaluate the performance of NormNet, the Peterson and Barney (1952) database (Peterson and Barney, 1952; Watrous, 1991) was chosen because it has been widely used as a benchmark database for studying vowel identification. Peterson and Barney originally tape recorded 76 speakers (33 males, 28 females, and 15 children) each speaking 10 vowels twice in a /h**V**d/ context, resulting in 1,520 tokens. The vowels used are found in Table 1. The recorded vowels were analyzed and the steady state measurements for F0, F1, F2, and F3 were preserved in the dataset. Listeners in this original study achieved 94% accuracy in recognition tasks when evaluating these vowels in /h**V**d/ context.

Number	ARPAbet symbol	IPA Symbol	/hVd/
1	IY	i	Heed
2	IH	ɪ	Hid
3	EH	ɛ	Head
4	AE	æ	Had
5	AH	ʌ	Hud
6	AA	ɑ	Hod
7	AO	ɔ	Hawed
8	UH	ʊ	Hood
9	UW	u	Who'd
10	ER	ɜ	Heard

TABLE I. The ten vowels in the Peterson and Barney (1952) database.

Hillenbrand and Gayvert (1993) synthesized steady-state values corresponding to the values of the formants in the database in order to determine how well listeners can identify vowels based on static spectral cues. Seventeen listeners achieved 72.7% accuracy for the synthesized vowels with flat F0 contours, which hold F0 constant for the duration of the sound stimulus. When F0 movement was added, performance only slightly improved to 74.8% correct. For the purposes of the simulations in this paper, the Hillenbrand and Gayvert (1993) performance will be used as a basis of comparison and the methods of these simulations will attempt to adhere to the methods presented in that paper.

B. Procedure

A vowel synthesizer (Slaney, 1998) was used to generate steady-state versions of all 1,520 tokens in the Peterson and Barney (1952) database. Formant frequencies and F0 were held constant for the full duration of the stimulus, similar to the synthesized vowels used by Hillenbrand and Gayvert (1993). The sampling frequency was set at 16 kHz and the formant bandwidth was set at 50 Hz. Waveforms for a sample synthesized vowel, 'IY' for a man, woman, and child are shown in Figure 7.

In order to assess the performance of the system, several types of simulations were performed. Simulations were conducted by varying vowel lengths, the dynamic range and number of filters of the filterbank, the training set size, and spectral inputs with and without F0 information combined in the mappings.

In order to simulate the natural variances across human listeners, the dynamic range of the filterbank and the number of filters were varied for the simulations. The inputs were presented to the model in random order. The simulations were run on a workstation PC using a dual core AMD Opteron Processor 246 with 1.99 GHz and 3.18 GB of RAM. Matlab v.7.1 and the auditory toolbox (Slaney, 1998) were used to run the simulations. Statistical analysis was conducted using Statistics To Use (Kirkman, 1996) and Wessa.net (Wessa, 2007) software.

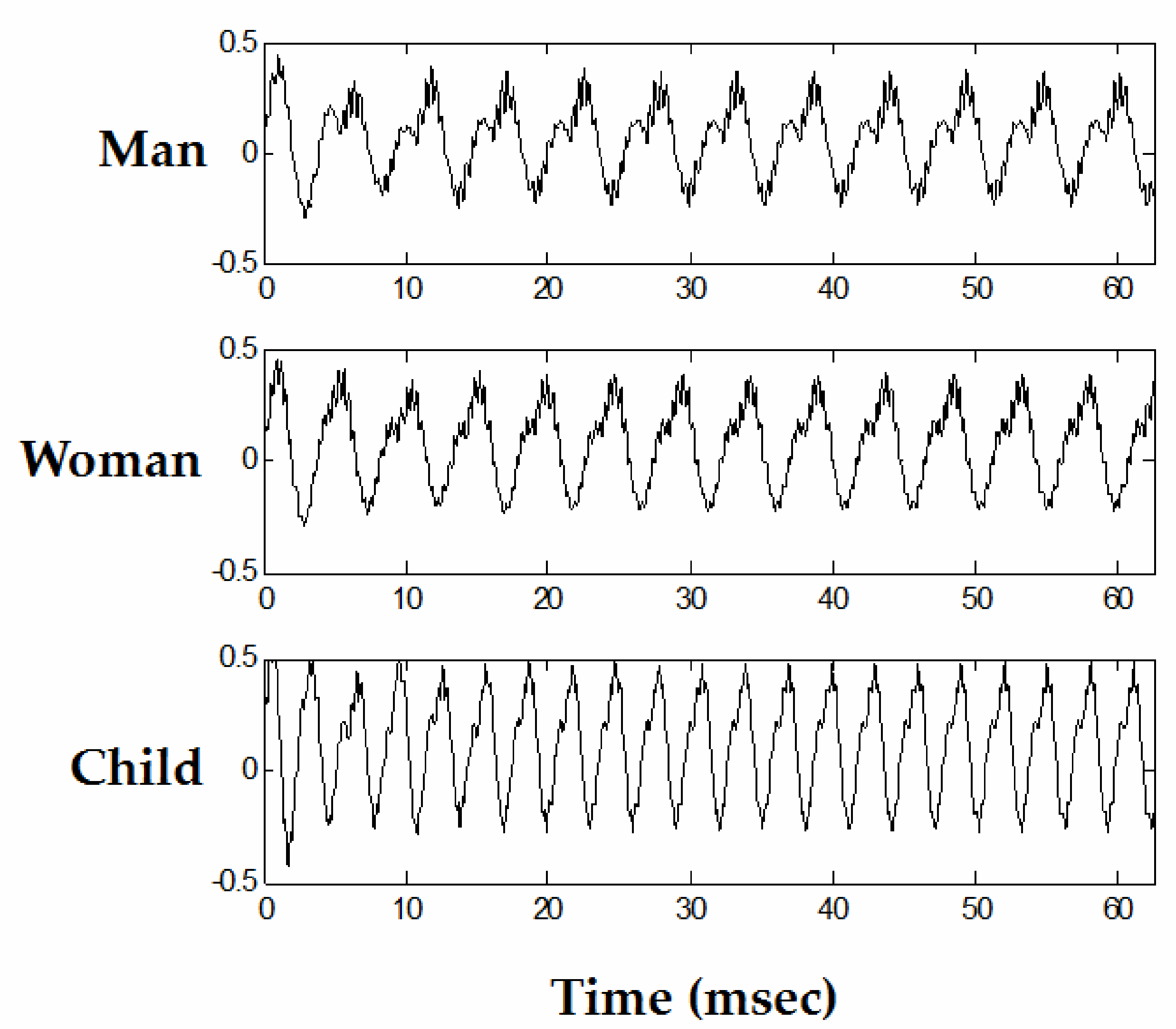

Figure 7: *Waveforms for synthesized steady-state vowel 'IY'.* The top plot corresponds to a male, the middle to a female, and the bottom to a child.

VI. RESULTS

A. Number of filters for the filterbank

Simulations were performed by varying the number of filters (100, 150, 200, 250, 300, 350, 400) while keeping the filter range constant at 50-7500 Hz, 400 tokens in the training set, and three runs for each filterbank size. These simulations were performed both with and without adding F0 information to the input-activated spectral information in the Anchor Map. The results from these simulations are illustrated in Figure 8. Interestingly, adding F0 caused model performance to deteriorate by approximately 5%. An analysis of variance (ANOVA) did not show a significant effect for filterbank size (F[6,14]=0.338, p<0.91).

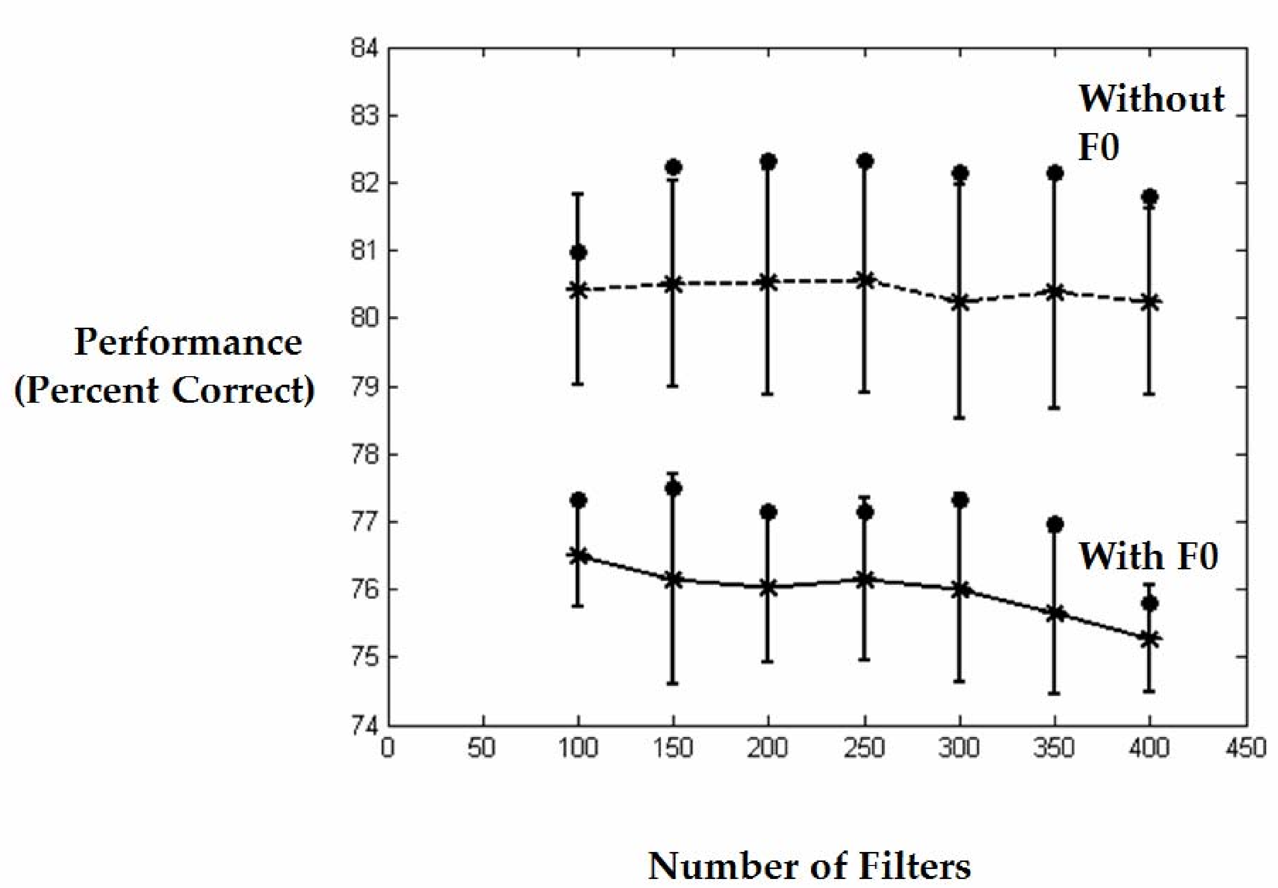

Figure 8: *Filterbank size.* The filterbank size was varied from 100-400 filters. The dataset was tested both with and without F0 information added to the spectral information of the Anchor Map. Three runs were performed with each filterbank size. The dynamic range was from 50-7500 Hz and the training set size was set to 400 tokens. The dashed line shows the results without F0 information and the solid line shows the results with F0 information in the spectral map. The * indicates the mean plus error bars and the • indicates the best performance for that filterbank size.

B. Dynamic range of filterbank

The dynamic range of the filterbank was tested; see Figure 9. The low frequency was varied from 20-180 Hz while the high end was held constant at 7500 Hz. The results of these simulations are illustrated in Figure 9a.

These simulations were performed without adding F0 information, with 250 filters, and with a training set size of 400 vowel tokens. The performance of the system was best when the low frequency end was below 100 Hz (approximately 80% correct on average). These data were found to be well fit by a linear model with a negative slope (for the low frequency value: mean performance $R^2 = 0.87$ and best performance $R^2 = 0.94$). Performance deteriorated gradually as the lowest frequency was increased from 20 Hz because some of the lower frequency vowels in the databank contain frequency information below 100 Hz.

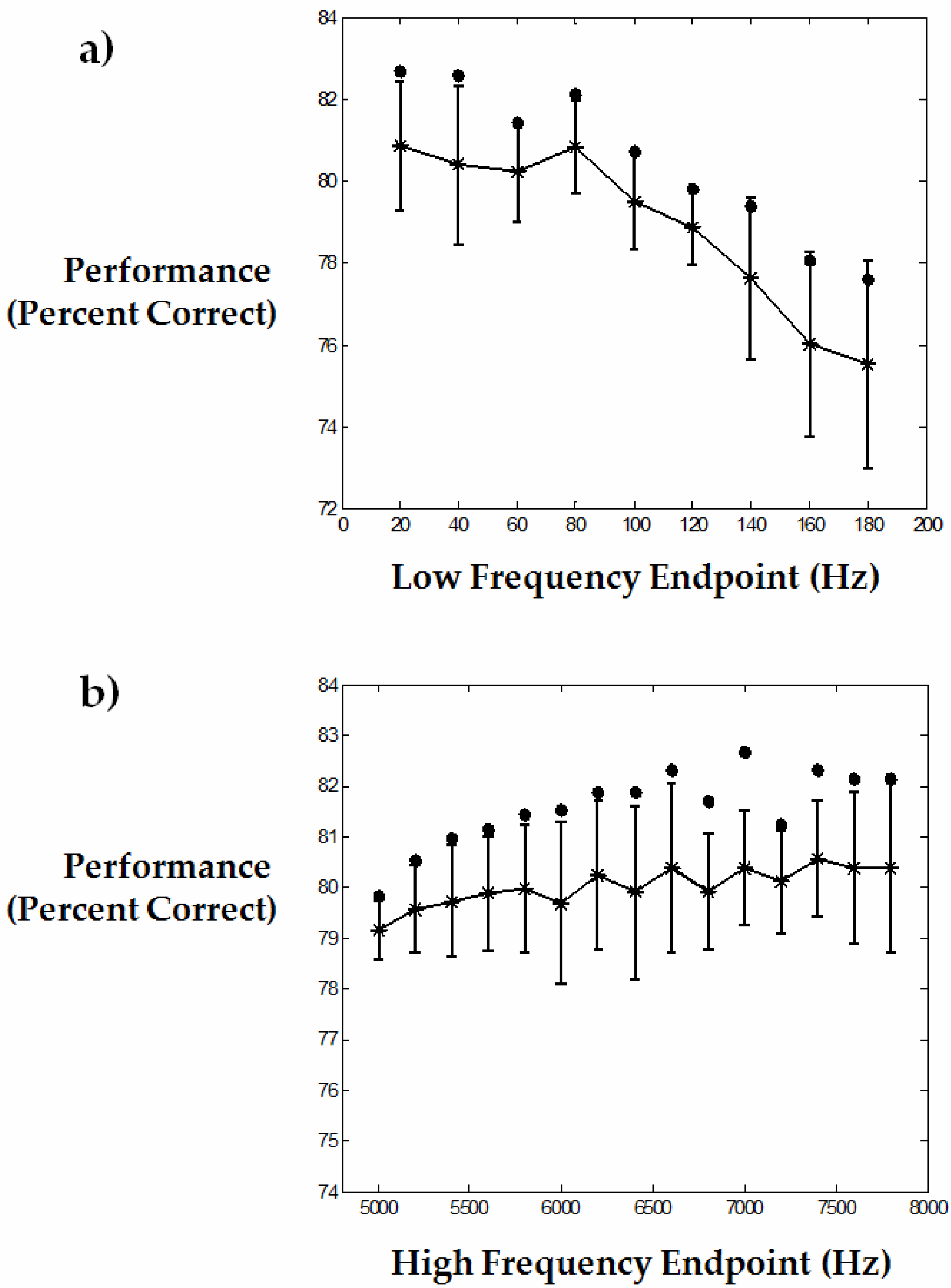

Figure 9: *Dynamic range of filterbank.* 250 filters were used in the filterbank, the training set contained 400 tokens, and three runs were performed for each variation. No F0 information was added to the spectrum for these simulations. The * indicates the mean plus error bars and the • indicates the best performance for that variation in the filterbank. **(a)** The low frequency endpoint was varied from 20-180 Hz while the high frequency endpoint was held constant at 7500 Hz. **(b)** The high frequency endpoint was varied from 5-8 kHz while the low frequency endpoint was held constant at 50 Hz.

The high frequency was also varied from 5-8 kHz while the low end was held constant at 50 Hz. The results of these simulations are found in Figure 9b. The high frequency manipulation caused less of an effect. When these data were fit to a linear model (for the high frequency value: mean performance $R^2 = 0.722$ and best performance $R^2 = 0.629$), the slope was nearly zero (mean performance = 0.00036 and best performance = 0.00067) indicating that there is little change in performance across the different high frequency endpoint values. This is because the high frequency range takes into account information above the F3 values and the vowels used in these simulations were synthesized with only F0-F3 information. Generally, the effect of manipulation of the higher formants of the vowel (F3-F5) has a much a smaller effect on vowel identification (Johnson, 2005; Slawson, 1968; Nearey, 1989) and thus we expect that it would have less effect on the model even if we were to test items that contained such information. This prediction has yet to be tested, however.

C. Training set size

Training set size was varied (100, 200, 300, 400, 500, 600) with and without adding F0 information contained in the Anchor Map. The training set was chosen randomly without replacement. The remainder of the dataset was used for testing. The simulations used a filterbank that consisted of 250 filters ranging from 50-7500 Hz. Figure 10 shows the overall performance results from the different training set sizes. When the training set contained only 100 tokens, performance was the worst near 73% correct. Based on these results, a training set size of 400 achieves the best performance without adding F0 information (82.23 % correct). Again, the model performed better without adding F0 information in order to anchor the spectral map.

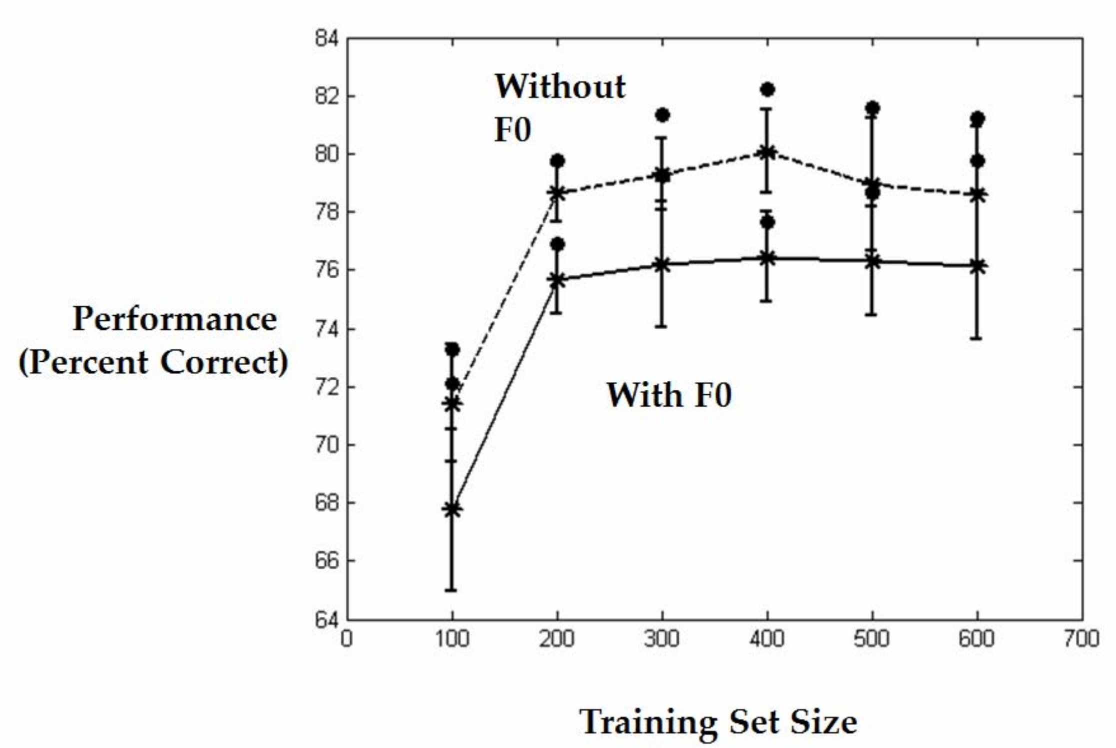

Figure 10: *Training set size.* The overall performance when the training set size is varied from 100 to 600 vowels. The dataset is also tested both with and without F0 information

added to the spectral information of the signal at the Anchor Map. Five runs were performed for each training set size. The filterbank consisted of 250 filters ranging from 50-7500 Hz. The best performance, mean, and standard deviation results were recorded. The dashed line shows the results without F0 information added and the solid line shows the results with F0 information added to the spectral map. The * indicates the mean plus error bars and the • indicates the best performance for that training set size.

D. Vowel duration

Vowel duration was varied to determine if steady state vowel duration affects model performance. Three vowel durations were tested (62.5, 300, 600 msec). No additional F0 information was used. The filterbank consisted of 240-260 filters, the dynamic range varied from 20-100 Hz on the low end and 7-8 kHz on the high end. The training set size was held constant at 400 vowel tokens. Four simulations were run for each vowel duration. Table 2 shows that varying vowel duration had little effect (F[2,9] = 0.2428, $p < 0.8$). However, the small effect may be because these vowels are steady-state. Varying the duration of naturally-produced vowels there may lead to a different outcome.

Vowel Duration(msec)	**Best Performance**	**Mean**	**Standard Deviation**
62.5	81.61	79.96	0.81
300	79.91	79.28	0.49
600	80.80	79.96	0.87

TABLE II. The overall performance when the vowel duration varied from 62.5msec, 300 msec, and 600 msec. The filterbank varied from 240-260 filters, and the dynamic range varied from 10-100 Hz at the low end and 7-8 kHz at the high end. The training set size contained 400 tokens and four runs at each vowel duration were tested. No F0 information was added to the spectrum for these simulations. Percent correction classification in terms of best performance, mean, and standard deviation are recorded.

E. With and without adding F0 information

Three types of F0 simulations were performed. The first did not include any additional F0 information. In the second set, F0 information was only added to the Anchor Map and the Spectral Map received only spectral information. In the last type of simulations, F0 information was added to the spectral representation and this combination was redundantly mapped as input into both the Anchor Map and Spectral Map. F0 information was added to the spectrum to increase the energy at the filter corresponding to F0. These simulations used a training set of 400 vowel tokens of 62.5 msec in duration, a filterbank of 240-260 filters, a low frequency from 20-100 Hz, a high frequency of 7-8kHz, and 14 runs of each simulation type. Table 3 summarizes the results of these simulations. The best performance of 81.61% was found with no F0 information added. The simulations did find a highly significant effect across these conditions (F[2,39] = 41.58, $p < 0.001$) indicating that adding F0 information impaired performance.

F0 information	**Best Performance**	**Mean**	**Standard Deviation**
Without F0	81.61	79.96	0.81
Only in the Anchor Map	77.68	76.45	1.54
In both the Anchor and Stream Maps	78.04	77.33	0.64

TABLE III. The overall performance without F0 information added to either map, F0 information only added to the Anchor Map, and F0 information added to the spectrum and redundantly mapped into both the Anchor Map and the Stream Map. The filterbank varied from 240-260 filters, and the dynamic range varied from 10-100 Hz at the low end and 7-8 kHz at the high end. The training set contained 400 tokens. The vowel duration was set at 62.5 msec. Fourteen runs were performed for each simulation type. Percent correction classification in terms of best performance, mean, and standard deviation are recorded.

VII. DISCUSSION

A. Comparison to human listeners

Comparisons of human identification rates and those of the model can only hope to show a qualitative correspondence, because human speakers come to such experiments with full knowledge of a language, and may thus contextually process even reduced speech cues in a way that a model that learns only those cues cannot. In particular, humans may experience filtering and competitive interference properties that a simple model may not. Despite this caveat, human/model comparisons illustrate many similar properties, as noted below.

Simulated identification rates for the synthesized vowels are shown in Table 4a along with the identification rates reported by Hillenbrand and Gayvert (1993) for flat F0 stimuli. Table 4b shows the results broken down across speaker groups (men, women, and children) from Hillenbrand and Gayvert (1993) and the model. The identification rates across speaker groups for the simulations was significant (F[2,39] = 147.1105, $p < 0.001$). The error rate was lowest for child speakers for both humans and the model. The model had the lowest error rate for the women speakers and the human listeners performed the best for male speakers. This difference may be due to learning in which the model was exposed to randomly chosen synthesized vowel samples from all three speakers groups whereas human listeners are exposed to a much wider variety of samples in varying context.

The confusion matrices for the Hillenbrand and Gayvert (1993) study are shown in Table 5a and for the simulations in Table 5b. The simulations performed in this study found an overall accuracy measure of 79.96%, which is better than the 72.7% reported by Hillenbrand and Gayvert (1993).

Vowel	Simulation results	Hillenbrand and Gayvert (1993) Flat F0
IY	95.53 ± 1.10	96.2
IH	86.90 ± 4.91	67.0
EH	70.28 ± 3.13	65.8
AE	81.98 ± 1.53	63.2
AH	83.96 ± 3.81	74.7
AA	80.39 ± 1.83	55.0
AO	70.99 ± 7.03	67.2
UH	81.35 ± 1.97	62.0
UW	66.10 ± 3.84	89.1
ER	82.03 ± 4.11	86.6
TOTAL:	**79.96 ± 0.81**	**72.7**

(a)

Talker group	Simulation results	Hillenbrand and Gayvert (1993) Flat F0
Men	80.35 ± 1.78	74.4
Women	82.76 ± 0.54	72.2
Children	72.71 ± 2.10	70.0

(b)

TABLE IV. (a) Percent correct identification rates for the flat-formant synthesized vowels and **(b)** Percent correct identification rates for the three talker groups (men, women, and children) with flat F0s in both the simulations performed in this study and in the Hillenbrand and Gayvert (1993) study.

The confusion matrices for both the Hillenbrand and Gayvert (1993) study and the simulations reported here show that most errors occurred near the diagonal. The layout of the confusion matrix roughly corresponds to the layout of the vowels in F1/F2 space; see Figure 11. F1/F2 space is considered a rough perceptual mapping of vowels in that there is a relationship between the intended vowel and the formant frequency pattern (Peterson and Barney, 1952). Figure 11a shows the vowels that were classified correctly and Figure 11b shows the vowels that were classified incorrectly. The ellipses are drawn based on the Peterson and Barney (1952) dataset. In Figure 11b, it is apparent that the majority of the vowel classification errors are near misses which occurred on vowel boundaries in the perceptual F1/F2 space. Fuzzy ARTMAP did a good job of correctly classifying vowels in overlapping F1/F2 space because the system does not cluster vowel boundaries in only F1/F2 space. Rather it takes into account the entire normalized spectra of the vowels.

	IY	**IH**	**EH**	**AE**	**AH**	**AA**	**AO**	**UH**	**UW**	**ER**
IY	96.2	3.1	0.6	0	0	0	0	0	0	0
IH	25.1	67.0	6.7	0.3	0.1	0	0	0.6	0	0.1
EH	1.3	23.7	65.8	7.2	0.3	0	0	0.4	0.1	1.1
AE	0.1	0.6	28.0	63.2	2.0	4.0	0	0.3	0	1.9
AH	0	0.1	0.9	0.7	74.7	12.8	6.8	2.7	0.1	1.2
AA	0	0	0.2	0.1	13.6	55.0	30.5	0.6	0.1	0
AO	0	0	0	0	8	5.9	67.2	13	5.9	0
UH	0	0.2	0.1	0	5.2	0.1	3.1	62.0	28.4	0.9
UW	0.2	0.2	0	0	0.7	0	0.7	9	89.1	0.2
ER	0.3	4.1	4.0	0.3	0.9	0	0	3	0.7	86.6

(a)

	IY	**IH**	**EH**	**AE**	**AH**	**AA**	**AO**	**UH**	**UW**	**ER**
IY	95.53	4.47	0	0	0	0	0	0	0	0
IH	7.24	86.90	4.28	0	0	0	0	0	0	1.58
EH	0	16.79	70.28	7.03	0	0	0	0	0	5.90
AE	0	0	9.25	81.98	7.76	0.31	0	0.12	0	0.56
AH	0	0	0	2.22	83.96	10.26	3.07	0.25	0	0.25
AA	0	0	0	2.40	13.50	80.39	2.90	0.81	0	0
AO	0	0	0	0.33	14.56	8.49	70.99	4.62	1.01	0
UH	0	0	0	0	0.90	0	4.48	81.35	9.02	4.25
UW	0	0.12	0.31	0	0	0	1.73	29.21	66.10	2.53
ER	0	2.72	3.06	3.54	2.72	0	0	5.53	0.41	82.03

(b)

TABLE V. (a) Confusion matrix reported by Hillenbrand and Gayvert (1993) for synthesized steady state vowels with flat F0 contours. **(b)** Confusion matrix for synthesized steady state vowels with flat F0 contours generated with the model. The simulation results reported here are the mean results from fourteen runs. In each run, the filterbank was randomly chosen to be from 240-260 filters, and the dynamic range from 10-100 Hz at the low end and 7-8 kHz at the high end. The training set contained 400 tokens. The vowel duration was set at 62.5 msec.

Hillenbrand and Gayvert (1993) found that confusions between ‘IY’ and ‘IH’, where ‘IH’ was heard as ‘IY’; between ‘UH’ and ‘UW’, where ‘UH’ was heard as ‘UW’; and between ‘EH’ and ‘AE’ where ‘AE’ was heard as ‘EH’ are tense-lax asymmetries. They hypothesized that, when the subjects listened to vowel stimuli without durational cues, they had a tendency to misclassify the vowels as long rather than short vowels. The ‘IH’ and ‘IY’ confusions were consistently encountered with the model. However, the model was not as susceptible to the ‘EH’ and ‘AE’ confusions. This may be due to the shorter vowel duration (62.5 msec) used by the model as compared to the longer duration (300 msec) used by Hillenbrand and Gayvert (1993) such that the short ‘AE’ was not misclassified as the long ‘EH’ when presented with shorter stimuli.

In the simulations with differing vowel durations these confusions were reversed with ‘UW’ heard as ‘UH’. These differences may be due to the initial stimulus set-up. When an analysis of variance (ANOVA) is performed at the vowel durations of 62.5 msec, 300 msec, and 600 msec, it was found that the correct classification of 'UH' differs significantly across the different durations (F[2,9] = 4.315, $p < 0.05$) and almost significantly for the correct classification of 'UW' (F[2,9] = 3.320, $p < 0.09$). The best classification was found at the 300 msec vowel duration, which is the same duration used by Hillenbrand and Gayvert (1993). Thus, the 300 msec stimuli seems to provide the best performance for classification of 'UH' where the human subjects had a tendency to classify these vowels as long. When the model was presented with much shorter stimuli (62.5 msec), it classified these vowels as short, with performance improving at the longer vowel durations.

One other difference between Hillenbrand and Gayvert (1993) and the simulations concerns the classification of the vowel 'AA'. Hillenbrand and Gayvert (1993) found that human listeners frequently misclassified 'AA' as 'AO' whereas the model frequently misclassified 'AA' as 'AH', 'AH' as 'AA' and 'AO' as 'AH'. All three of these vowels are back vowels for which the tongue is placed near the back of the mouth and roughly corresponds to a smaller difference between F1 and F2 (Lindau, 1978). 'AA' is differentiated from 'AH' and 'AO' in that it is slightly more open and unround, with 'AH' unrounded and 'AO' rounded. Thus, 'AA' and 'AH' differ only by a slight variation in openness, whereas 'AA' and 'AO' also differ in rounding. Finally, Figure 11 shows that all three of these vowels significantly overlap in F1/F2 space. The confusions made by both human listeners and the model are consistent with the close proximity of these vowels in perceptual space and that both types of confusions are valid.

The last group of differences involves taking into consideration the better classification of 'AE' and 'EH' by the model versus the better classification of 'IY' by human listeners. It seems that the model is biased towards lower values for F1 than human listeners. This may be due to the fact that the low frequency endpoint of the filterbank was varied from 10-100 Hz, which may be lower than what is typically found in humans. These differences may also be attributed to the training effects where the model had more exposure to lower harmonics than do humans when learning to speak. In making this comparison, it is also worth noting that the model experienced only the vowel data, whereas humans respond with potential competition from the entire language.

The classification results show that the speaker normalization circuit helped to recognize vowels in the Peterson and Barney (1952) dataset. The training results of the fuzzy ARTMAP classifier also produce important information. During training, fuzzy ARTMAP learns categories corresponding to the vowel categories. In these simulations, only ten categories were learned, corresponding to the ten vowels in the dataset. Therefore, there is a one-to-one mapping between the learned categories and the vowel categories, which indicates the success of the system in creating invariant representations of the vowel categories. If the normalization scheme did not perform well, fuzzy ARTMAP would have learned to select many more categories corresponding to each vowel category. For example, without speaker normalization pre-processing, the classifier generated learned, on average, thirty categories and the vowel classification performance dropped to 71.95%. The classification rate of the system without normalization was slightly lower than human performance of 72.7% as reported by Hillenbrand Gayvert (1993).

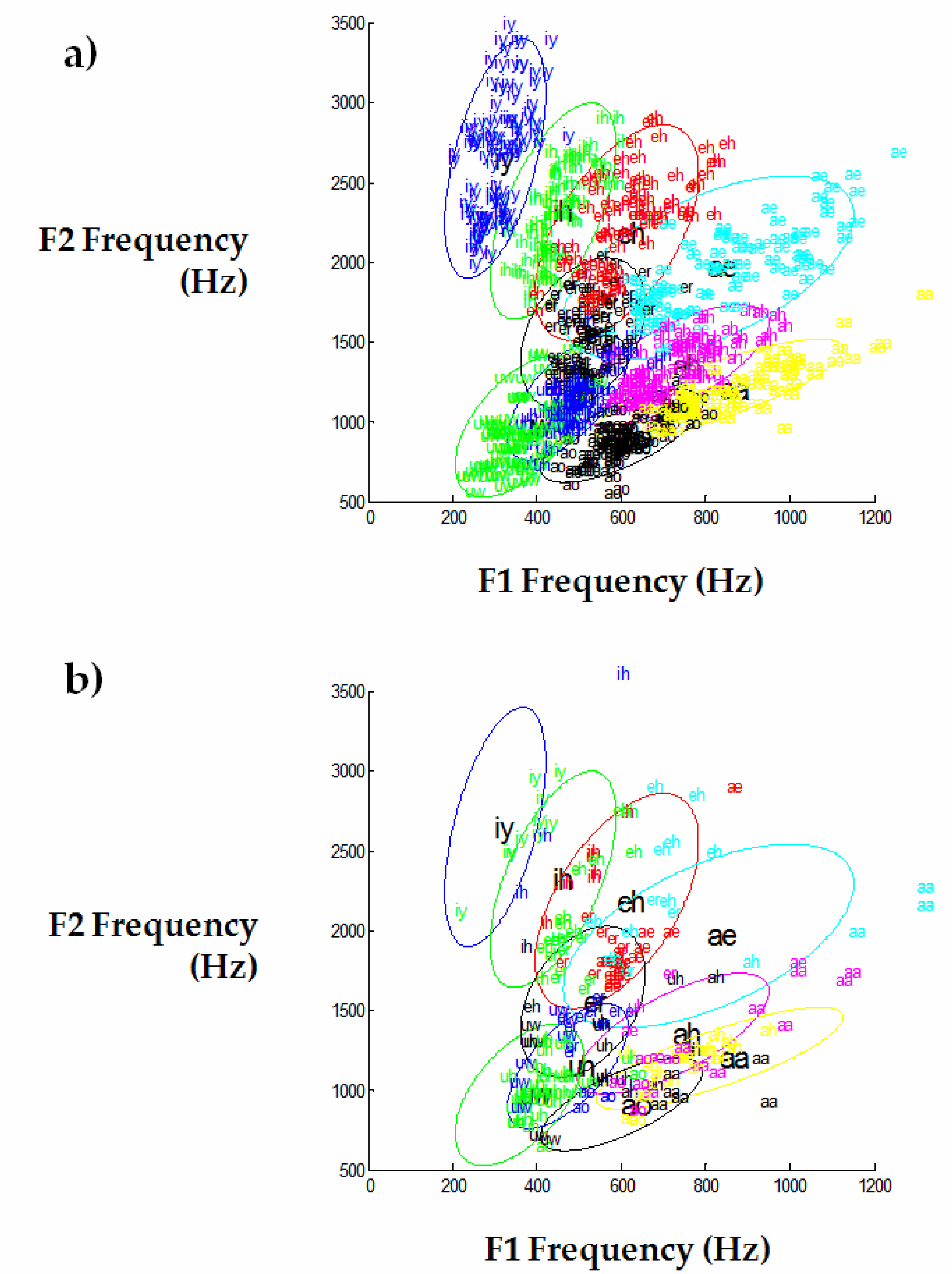

Figure 11: *F1/F2 vowel space*. The ellipses correspond to the confidence intervals reported over the entire Peterson and Barney (1952) database where 'IY' is blue, 'IH" is green, 'EH' is red, 'AE' is light blue, 'AH' is pink, 'AA' is yellow, 'AO' is black, 'UH' is dark blue, 'UW' is green, and 'ER' is black. **(a)** The individual data points correspond to the locations of the correctly classified vowels in the simulations. **(b)** The individual data points correspond to the misclassified vowels. The color of the data points corresponds to what the vowel should have been classified as and the vowel label at the data point corresponds to what it what misclassified as.

As summarized above, the vowel identification rates of the model are comparable to those reported by Hillenbrand and Gayvert (1993), with most of the misclassifications lying in adjacent F1/F2 space. The identification rate of 72.7% of Hillenbrand and Gayvert (1993) and the 79.96% reported in these simulations is still significantly less than the 94% accuracy reported by Peterson and Barney (1952) in a vowel recognition task because of the use of 3-formant steady state synthesized vowels. These stimuli are more difficult for listeners and the model because of their lack of context, durational variation, and natural fluctuations of speech. Accuracy greatly improves from 57%-95% for isolated vowel recognition to 83%-96% in /cVc/ context (Nearey, 1989). Listeners achieved 79% accuracy when listening to vowel recordings, but when the vowels were synthesized with a fixed duration and steady-state formants, only 61% accuracy was achieved (Lehiste and Meltzer, 1973). Listeners could achieve 89% accuracy when vowels were synthesized with their original formant trajectories, but would only achieve 74% accuracy when the vowels were synthesized with flat formants (Hillenbrand and Nearey, 1999).

The ART modeling framework, of which the NormNet forms a part, promises to generalize to more complex speech. Variability in human speech such as durational cues and context cues are best captured by a dynamical system that catches sources of variation in time, that is able to learn from local knowledge available in the speech signal, and that learns in real time. Previous modeling studies have quantitatively simulated speech categorization and word recognition data with these properties (e.g., Grossberg *et al.*, 1997; Boardman et al., 1999; Grossberg and Myers, 2000). The NormNet circuit, when integrated within this emerging theory, will enable it to begin to simulate variable speech perception properties of multiple speakers.

Many studies have shown that speaker normalization is an active process such that listeners are better and faster at word and vowel recognition in single talker lists than in multi-talker lists (Creelman, 1957; Summerfield and Haggard, 1975; Verbrugge *et al.*, 1976) and that listeners retain memory of speaker-specific acoustic detail that influences their memory of previously spoken words (Palmeri *et al.*, 1993; Church and Schacter, 1994; Goldinger, 1996, 1997). Although some authors have suggested that this is evidence against automatic speaker normalization, these results are consistent with the NormNet model. The increased reaction time and decreased performance in the multi-speaker lists can be attributed to the switching cost of moving to a new stream and to a new anchor and location within the maps. More generally, such effects may be due to a range of additional interactions between speaker-dependent and speaker-independent processes. This hypothesis is further supported by the study performed by Kato and Kakehi (1988), which showed that accuracy in syllable recognition monotonically increases from the first to the fifth presentation and that accuracy no longer increases on successive presentations. Furthermore, listeners are influenced by expectation of gender either through auditory cues (Eklund and Traunmüller, 1997; Johnson *et al.*, 1999) or visual cues (Walker *et al.*, 1995; Strand and Johnson, 1996; Schwippert and Benoit, 1997; Johnson *et al.*, 1999). Taken together, these studies may probe how the Anchored Stream serves as a frame of reference for understanding a speaker and that mismatched expectations or switching between speakers affects reaction time and performance as the model adapts to the variation in the location of the Anchored Stream.

Kraljic and Samuel (2007) recently reported that although listeners may adjust their internal representations of phonemic categories based on the speaker, this is not always the case. In fact, some of these categorical adjustments are not related to the speaker but may be primed by other cues (e.g., rate in VOT of stop consonants). Therefore, it is necessary for the phonemic categories to learn in real time to adjust to fluctuations both within the speech of a single speaker and across speakers. Such adaptations can occur at multiple levels of the speech perception system.

Johnson (1997a, 1997b, 2005, 2006) suggests that the above-mentioned evidence is supportive of an episodic/exemplar coding model that bases speech perception on a set of stored exemplars that adapt to the perceived identity of the talker. The exemplar models do not make use of speaker normalization and instead rely on the large repertoire of exemplars to feed into category nodes describing the recognized speech sample. However, the limited data simulated with these models fails to show how they can scale up to natural speech and the large variety of speakers encountered in everyday life. To assume that the exemplars are created for each different mode of speaking the same speech sound would assume that the human brain has a massive capacity to store these exemplars, and a way to search among these exemplars in real time. The simulations reported by Johnson (1997b, 2006) do not report the number of exemplars or categories that are learned for each of the words tested. Furthermore, the learning of the word categories based on the set of exemplars requires the use of teacher-based knowledge to create the mapping between the exemplars and the category nodes. The category nodes cannot be created based on local knowledge and in real time.

NormNet is not subject to these sources of criticism. The speaker normalization circuit does not require the storage of instances or exemplars of speech sounds and can learn a single categorical representation for each speech sound. The creation of the invariant speech representations described herein does not require the use of a teaching signal and can be performed in real time with only local knowledge available to the system. The simulation results without normalization show that vowel categorization was poorer (71.95%) and that more vowel categories were created (30.29 nodes generated). Hence, the addition of speaker invariance creates more stable categories and better performance.

More generally, ART-based categorization can, where task demands require, learn both specific, concrete categories, even individual exemplars, as well as general, abstract categories. This learning process enables ART models to selectively pay attention to the learned prototype of critical feature patterns that predict successful performance. High vigilance leads to concrete categories, whereas low vigilance leads to general categories (see Section IV and the Appendix). Thus, one does not need to store all exemplars to learn fine distinctions when, in fact, a language requires them. Thus, in several senses, the NormNet circuit and the larger ART speech perception network to which it belongs (Figure 1), embody the phonological principle which states “that languages have basic building blocks, which are not meaningful in themselves, but which combine in different ways to make meaningful forms” (Pierrehumbert, 2006).

B. Comparison to other speaker normalization techniques

Other speaker normalization techniques have been applied to a variety of vowel recognition tasks using different classifiers. Two cues, vocal tract length and F0, are important in these speaker normalization techniques. Inter-speaker variability is often attributed to the difference in the shape and length of the vocal tracts, with males typically having longer vocal tracts then females (Lee and Rose, 1998; Stevens, 1998). The correlation between the vocal tract length and the position of the vowel formants contributes to differences perceived by the listener (Fant, 1973). Vocal tract length normalization (VTLN) is based on the assumption that the speech spectrum of one speaker differs from another due to stretching or compression along the frequency axis (Eide and Gish, 1996; Wegman *et al.*, 1996; Lee and Rose, 1996, 1998; McDonough and Byrne, 1999;

Dognin and El-Jaroudi, 2003; Glavitsch, 2003). The speech sound is normalized by warping the frequency axis onto a standard vocal tract length. It is, however, unclear how speakers could estimate vocal tract length during naturally occurring language experiences.

Nonlinear (e.g. Eide and Gish, 1996), linear (e.g. Zahorian and Jagharghi, 1991), and bilinear (e.g. Glavitsch, 2003) transformations have all been used in VTLN techniques. The resulting transformation has the same Fourier transform as the original except that it is warped along the frequency axis. Both linear and bilinear transformations have led to increased performance in systems performing speech related tasks (Lee and Rose, 1996; Wegmann *et al.*, 1996; Zhan and Westphal, 1997; Zhan and Waibel, 1997).

Wegmann *et al.* (1996) used a VTLN method in which the frequency warping was done using a piecewise linear transformation of the frequency axis with fixed points at 0 kHz and the Nyquist frequency. Ten warp scales were constructed and each map scale was applied to the speech sound. The best warp scale was chosen through a comparison to a generic voiced speech model. Wegmann *et al.* (1996) reported a 12% reduction in word error rate as compared to unnormalized gender-independent models and a 6% reduction as compared to unnormalized gender-dependent models when tested on the standard Switchboard Corpus (NIST).

Zahorian and Jagharghi (1991) evaluated the effect of both a linear transformation of spectral features and a speaker-dependent frequency warping procedure to evaluate improvement on vowel classification. In both, the normalization parameters were chosen to minimize the mean squared error between the normalized features and the target features. They found an 8-15% increase in accuracy, where the accuracy level ranged from 69-91%.

It is difficult to compare the performance across these different speaker normalization techniques because of the different data sets and vowel classifiers that were used. A meaningful metric is to compare the performance of each technique to human listeners on comparable tasks. The Peterson and Barney (1952) database contains only steady state vowel information and human listeners are not as good at recognizing steady-state vowels as vowels containing durational and contextual cues. Taking this into account, if the results from the simulations of this paper, 79.96% correct, are compared to the human listeners of the Hillenbrand and Gayvert (1993) study, 72.7% correct, the simulations reported by Nearey (1979), with 81%-92% correct, Sydral and Gopal (1986), with 81.8%-85.7% correct, and Turner and Patterson (2003), with 79-84%, it seems that these other systems may overfit human data, whereas the simulations from this paper adhere more closely to the reported human data.

C. Role of F0 in speaker normalization

F0 is determined by the rate of vibration of the vocal cords of the speaker and thus correlates with the size of the speaker's vocal folds (Titze, 1994). The average values of F0 are lowest in males, around 100 Hz, 200 Hz in females, and up to 400 Hz in infants (Kent and Read, 1992). Because the harmonics of the speech sound correspond to integer multiples of F0, F0 can, in principle, be inferred by the human brain from the spectrum of harmonics even if it is "a missing fundamental" in the signal (Pantev *et al.*, 1989; Ragot and Lepaul-Ercole, 1996).

The distance between F1 and F0 in critical bandwidth is an important cue for perception of vowel openness (Traunmüller, 1981). Speaker-dependent information contained in F0 has also been found to be important in both the recognition of vowels and Mandarin Chinese tones (Johnson, 1990; Moore and Jongman, 1997).

F0 varies greatly amongst speakers and the range of F0 in human speakers can vary from 50-800 Hz (Hess, 1983; Ferreira, 2007). The high end of this range, found in female and child speakers and in singing, can result in F0 being comparable to or higher than F1. In addition, it may be the case that the vowel's spectrum may contain a sufficiently small amount of energy for F0. In both of these situations, the asymmetric competition in the model may be compromised. In order to test this concern, additional F0 information added into the speaker normalization model to anchor the spectral information, which caused performance to decrease. Thus, the F0 information contained in the original vowel spectrum is sufficient for speaker normalization.

F0 has been found to be slightly helpful in understanding speech in both speech recognition systems and human listeners (Glavitsch, 2003; Magimai-Doss *et al.*, 2003). For example, Nearey *et al.* (1979) classified the Peterson and Barney (1952) database with a linear discriminant classifier to identify vowels. They reported 81% correct when the system was trained on log-transformed F1 and F2, 86% correct when F0 and F3 were included, and 92% correct when speaker mean log formant values were subtracted from the individual log formant values. Sydral and Gopal (1986) also used a linear discriminant classifier in which they achieved 81.8% correct when trained on F0 and F1-F3 and 85.7% correct when also trained on three bark-transformed spectral differences (F3-F2, F2-F1, and F1-F0). Turner and Patterson (2003) used a Mellin transform to look at the variation of vocal tract length and achieved 79-84% correct. These improvements using F0 are small or modest. In the case of Nearey *et al.* (1979), the classification used F0 information as an additional feature for the classifier rather than for use with the normalization scheme. Sydral and Gopal (1986) used F0 information only to normalize the first formant. Thus, although F0 contains important cues that are useful for human listeners to evaluate the meaning of a speech utterance, F0 may not be needed to normalize typical speech sounds. However, F0 may be more useful in normalizing more extreme cases such as singing or infant-directed maternal speech which in some cases has a very high F0. These cases have yet to be thoroughly tested in NormNet.

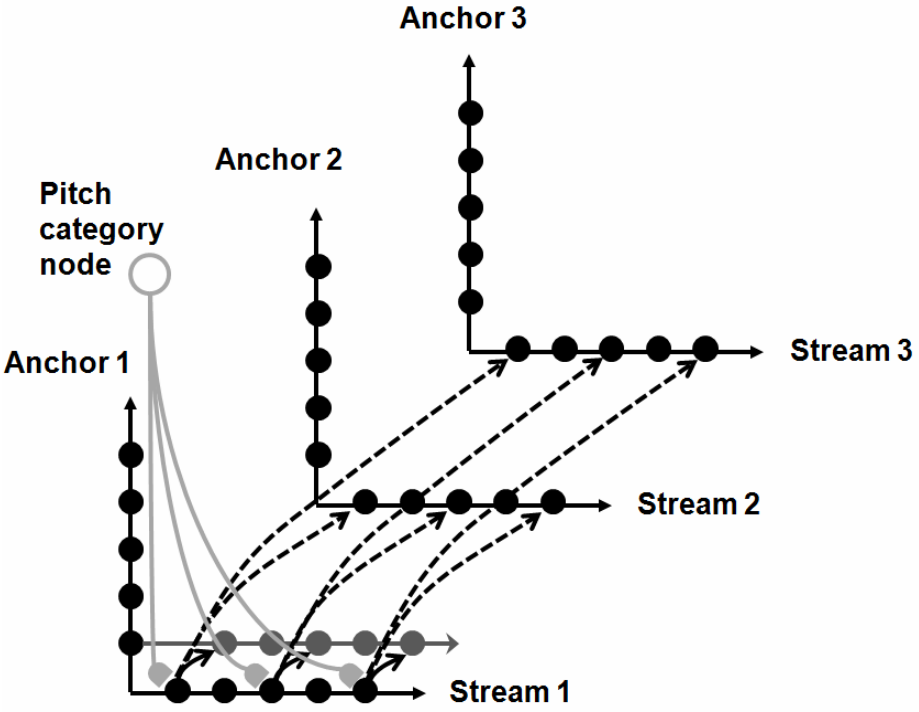

Figure 12: *Unification of multiple streams and their speaker normalization circuit.* Three potential streams and their anchor maps are illustrated. The first stream is chosen when its

pitch category wins a competition with other streams and uses a harmonic top-down expectation to select the frequencies that are compatible with that pitch. Asymmetric competition with the other streams causes the selected frequencies to be exclusively allocated to that stream. Other streams select their frequency spectra from the remaining frequencies. This selection process determines the anchor frequency for each stream and thereby initiates speaker normalization within each stream by again using asymmetric competition to normalize each selected frequency spectrum in its stream.

VIII. CONCLUSION

The NormNet model provides a proof of principle for a new insight into how speaker normalization may be carried out by the brain. The speaker normalization model was able to achieve accuracy of 79.96% correct, on average, which is comparable to the results obtained with human listeners identifying similar vowel stimuli. The model proposes that tonotopic strip maps of frequency-selective auditory cortical cells and asymmetric competitive interactions are used both to define the auditory streams that characterize acoustic sources, and to normalize the frequency spectra of these streams so that they can be understood across multiple speakers. Figure 12 depicts a hypothetical brain map that unifies multiple streams and the speaker normalization circuit. The way in which strip maps and asymmetric competition may be used in both streaming and speaker normalization circuits is a worthy topic of future research to clarify the predicted shared mechanisms that may be at work.

ACKNOWLEDGEMENTS
This work was supported in part by the National Science Foundation (NSF SBE-0354378) and the Office of Naval Research (ONR N00014-01-1-0624).

APPENDIX: FUZZY ARTMAP EQUATIONS

ARTMAP is a neural network that is capable of both unsupervised and supervised incremental learning in response to sequences of binary input vectors presented in real time (Carpenter *et al.*, 1991). Fuzzy ARTMAP can learn stable recognition categories in response to binary or analog input vectors (Carpenter *et al.*, 1992). Learning always converges because all adaptive weights are monotonically increasing.

The fuzzy ARTMAP system consists of two adaptive resonance theory modules, ART$_a$ and ART$_b$ that are linked together by an inter-ART module, F^{ab}, called a map field (see Figure 6). During supervised learning, both modules receive a stream of input patterns: $\{\mathbf{a}^{(p)}\}$ and $\{\mathbf{b}^{(p)}\}$ where $\mathbf{b}^{(p)}$ is the correct prediction given $\mathbf{a}^{(p)}$. The inputs to the ART modules are $\mathbf{A} = (\mathbf{a},\mathbf{a}^c)$ for ART$_a$ and $\mathbf{B} = (\mathbf{b},\mathbf{b}^c)$ for ART$_b$. These inputs are in a complement-coded form. Complement coding combines ON-cell and OFF-cell responses to prevent category proliferation by normalizing the amplitudes of the input feature vectors while preserving the amplitude of individual feature activations. To define complement coding, consider the ART$_\mathrm{a}$ module in which the input vector, **a**, is the ON-response. Then the complement of **a** is the OFF-response defined as:

$$\mathbf{a}_i^c = 1 - \mathbf{a}_i \, . \qquad (A1)$$

Hence, the complement coded input **A** is a 2*M* dimensional vector:

$$\mathbf{A} = (\mathbf{a},\mathbf{a}^c) = (a_1,...,a_M, a_1^c,...,a_M^c) \, . \qquad (A2)$$

Each ART module contains a field, F_0^a and F_0^b, of cells that represent a current input vector **a** and **b**, respectively. The F_1^a and F_1^b feature fields receive complement-coded inputs **A** and **B** from F_0^a and F_0^b, respectively, and top-down learned expectations from the F_2^a and F_2^b active learned categories. The number of cell populations in each field is arbitrary. For ART$_\mathrm{a}$, $\mathbf{x}^a = (x_1^a,...,x_{2M_a}^a)$ is the F_1^a output vector, $\mathbf{y}^a = (y_1^a,...,y_{N_a}^a)$ is the F_2^a output vector, and $\mathbf{w}_j^a = (w_{j1}^a, w_{j2}^a,..., w_{j2M_a}^a)$ is the *j*th ART$_a$ adaptive weight vector. For ART$_b$, $\mathbf{x}^b = (x_1^b,...,x_{2M_b}^b)$ is the F_1^b output vector, $\mathbf{y}^b = (y_1^b,...,y_{N_b}^b)$ is the F_2^b output vector, and $\mathbf{w}_k^b = (w_{k1}^b, w_{k2}^b,..., w_{k2M_b}^b)$ is the *k*th ART$_b$ adaptive weight vector. The adaptive weight vectors are associated with each F_2 category cell population j (j = 1,…, 2N_a) for ART$_a$ and k (k = 1,…, 2N_b) for ART$_b$. Each adaptive weight, or long-term memory (LTM) trace, of the weight vector is initially set to one indicating an uncommitted category. After the category is selected for coding, it becomes committed. In the present simulations, only the F_2^b field is implemented and its nodes are directly activated by category name labels.

The fuzzy ART module, ART$_a$, requires three parameters to be specified. These parameters are a choice parameter $\alpha > 0$, a learning rate parameter $\beta \in [0,1]$, and a vigilance parameter $\rho \in [0,1]$.

Category choice also occurs in both ART modules. The notation for the ART$_a$ module will be listed. The equations are the same for the ART$_b$ module except that the superscript is *b* and the *j* subscript in the F_2 field is *k*. For each input **A** and F_2^a node *j*, the choice, T_j, is defined by:

$$T_j(\mathbf{A}) = \frac{|\mathbf{A} \wedge \mathbf{w}_j|}{\alpha + |\mathbf{w}_j|} \, , \qquad (A3)$$

where the fuzzy and operator $\wedge$ is defined by:

$$(\mathbf{p} \wedge \mathbf{q})_i = \min(p_i, q_i)\,, \tag{A4}$$

and the norm | | is defined by:

$$|\mathbf{p}| = \sum_{i=1}^{M} |p_i|: \tag{A5}$$

for any M-dimensional vectors **p** and **q**.

The ART$_a$ module makes a category choice when at most one F_2^a cell population is active at a given time. This category choice is indexed by J:

$$T_J = \max\{T_j : j = 1, \ldots, N\}\,. \tag{A6}$$

If more than one T_j is maximal, then the category j with the smallest index is chosen. These cells become committed in order of j = 1, 2, 3, … When the Jth category is chosen, $y_J = 1$ and $y_j = 0$ for all $j \neq J$. The F_1^a activity vector, $\mathbf{x} = \mathbf{A}$ when F_2^a is inactive and $\mathbf{x} = \mathbf{A} \wedge \mathbf{w}_J^a$ if the Jth F_2^a node is chosen.

Resonance and reset are governed by the match value:

$$\frac{|\mathbf{A} \wedge \mathbf{w}_J^a|}{|\mathbf{A}|}\,. \tag{A7}$$

If *(A7)* is greater than or equal to the vigilance, ρ, then resonance occurs and learning ensues, as defined below. Otherwise, mismatch reset occurs, which results in the choice function T_J being set to zero for the duration of the input presentation to prevent persistent selection and learning of that category. A new index J is then chosen by *(A6)* and the search continues until a chosen J achieves resonance.

Resonance triggers learning such that, once the search ends, the chosen weight vector, $\mathbf{w}_J^a$ is updated:

$$\mathbf{w}_J^{a(new)} = \beta(\mathbf{A} \wedge \mathbf{w}_J^{a(old)}) + (1-\beta)\mathbf{w}_J^{a(old)}\,. \tag{A8}$$

Fast learning, as was used in the simulations in this paper, occurs when $\beta = 1$.

The map field, F^{ab}, links the two ART modules and is used to form predictive associations between ART$_a$ and ART$_b$ categories and to perform match tracking. The map field becomes active whenever one of the ART$_a$ or ART$_b$ categories is active, or when both are active only if ART$_a$ predicts the same category as ART$_b$ through the weights $\mathbf{w}_J^{ab}$. The output vector of the F^{ab} map field, $\mathbf{x}^{ab} = \mathbf{y}^b \wedge \mathbf{w}_J^{ab}$ if the Jth F_2^a category is active and F_2^b is active; $\mathbf{x}^{ab} = \mathbf{w}_J^{ab}$ if the Jth F_2^a category is active and F_2^b is inactive; $\mathbf{x}^{ab} = \mathbf{y}^b$ if F_2^a is inactive and F_2^b is active; and $\mathbf{x}^{ab} = \mathbf{0}$ if F_2^a is inactive and F_2^b is inactive. Thus, $\mathbf{x}^{ab} = \mathbf{0}$ when the prediction $\mathbf{w}_J^{ab}$ is disconfirmed by $\mathbf{y}^b$. This mismatch triggers an ART$_a$ memory search, or hypothesis testing, for a better match via match tracking.

During match tracking, the vigilance parameter of ART$_a$, ρ_a, increases in response to a predictive mismatch with ART$_b$ in order to ensure that predictive errors are not repeated on subsequent presentations of the input. The parameter ρ_a calibrates the minimum confidence that ART$_a$ must have in a recognition category activated by the input **A** in order for the ART$_a$ module to accept that category. Smaller values of ρ_a lead to broader generalization and higher code compression. By match tracking, the minimum amount of generalization necessary to correct a predictive error is sacrificed. In other words, the ARTMAP system embodies a minimax learning rule in which the system strives to minimize predictive error while maximizing predictive generalization.

At the start of the input presentation, ρ_a equals the baseline vigilance and the map field vigilance parameter is ρ_{ab}. If

$$| \mathbf{x}^{ab} | < \rho_{ab} \, | \mathbf{y}^{b} | , \qquad \textit{(A9)}$$

then ρ_a is increased until it is slightly larger than the match value in *(A7)*. Reset occurs and a memory search discovers the next ART$_a$ category to learn. With fast learning, the map field weights $\mathbf{w}_{jk}^{ab} = 1$ for all time when *J* learns to predict the ART$_b$ category name *K*.

REFERENCES

Ames, H.M. and Grossberg, S. (2006). "Neural dynamics of auditory streaming, speaker normalization, and speech categorization." *Soc. Neurosci. Abstracts*, Atlanta, GA.

Ames, H.M. and Grossberg, S. (2007). "Speaker normalization using cortical strip maps: A neural model for steady state vowel identification." *Comput. Cogn. Neurosci. Conf. Abstracts*, San Diego, CA.

Bendor, D. and Wang, X. (2005) "The neuronal representation of pitch in primate auditory cortex." *Nature* 436, 1161-1165.

Bendor, D. and Wang, X. (2006) "Neural representations of pitch in auditory cortex of humans and other primates." *Curr. Opin. Neurobiol.* 16, 391-399.

Bilecen, D., Scheffler, K., Schmid, N., Tschopp, K., and Seelig, J. (1998). "Tonotopic organization of the human auditory cortex as detected by BOLD-FMRI." *Hear. Res.* 126(1-2), 19-27.

Bladon, R.A., Henton, C.G., and Pickering, J.B. (1984) "Towards an auditory theory of speech normalization." *Lang. Commun.* 4(1), 59-69.

Boardman, I., Grossberg, S., Myers, C., and Cohen, M. (1999) "Neural dynamics of perceptual order and context effects for variable-rate speech syllables." *Percept. Psychophys.* 6, 1477-1500.

Bowers, J.S. (2002). "Challenging the widespread assumption that connectionism and distributed representations go hand-in-hand." *Cogn. Psychol.*, 45, 413-445.

Bradski, G. and Grossberg, S. (1995) "Fast learning VIEWNET architectures for recognizing 3-D objects from multiple 2-D views." *Neural Networks,* 8, 1053-1080.

Bregman, A.S. (1990) *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Bullock, D., Grossberg, S., and Guenther, F.H. (1993) "A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm." *J. Cog. Neurosci.*, 5, 408-435.

Cariani, P.A. and Delgutte, B. (1996a) "Neural correlates of the pitch of complex tones. I. Pitch and pitch salience." *J. Neurophysiol.* 76(3), 1698-716.

Cariani, P.A. and Delgutte, B. (1996b) "Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch." *J. Neurophysiol.* 76(3), 1717-34.

Carpenter, G.A. (1997) "Distributed learning, recognition, and prediction by ART and ARTMAP neural networks." *Neural Networks*, 10, 1473-1494.

Carpenter, G.A. and Grossberg, S. (1991) *Pattern recognition by self-organizing neural* *networks.* MIT Press, Cambridge, MA.

Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H., and Rosen, D.B. (1992) "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multi-dimensional maps." *IEEE Trans. Neural Networks* 3, 698-713.

Carpenter, G.A., Grossberg, S., and Reynolds, J.H. (1991) "ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network." *Neural Networks* 4, 565-588.

Cedolin, L. and Delgutte, B. (2005) "Pitch of complex tones: Rate-place and interspike interval representations in the auditory nerve." *J. Neurophysiol.* 94, 347-362.

Chey, J., Grossberg, S., and Mingolla, M. (1997) "Neural dynamics of motion grouping: From aperture ambiguity to object speed and direction." *J. Opt. Soc. Am.* 14, 2570-2594.

Church, B.A. and Schacter, D.L. (1994) "Perceptual specificity of auditory priming: implicit memory for voice intonation and fundamental frequency." *J. Exp. Psychol. Learn. Mem. Cogn.* 20(3), 521-533.

Cohen, M.A., and Grossberg, S. (1997) "Parallel auditory filtering by sustained and transient channels separates coarticulated vowels and consonants." *IEEE Trans. Speech Audio Process.* 5, 301-318.

Cohen, M.A., Grossberg, S. and Stork, D.G. (1988) "Speech perception and production by a self-organizing neural network." In Lee YC (ed) *Evolution, Learning, Cognition, and Advanced Architectures*. Hong Kong: World Scientific, 217-231.

Cohen MA, Grossberg S and Wyse L (1995) "A spectral network model of pitch perception." *J. Acoust. Soc. Am.* 98, 862-879.

Creelman, C.D. (1957) "Case of the unknown talker." *J. Acoust. Soc. Am.* 29, 655.

Desimone, R. (1998). "Visual attention mediated by biased competition in extrastriate visual cortex." *Phil. Trans. Royal Soc.*, 353, 1245-1255.

Dognin, P.L. and El-Jaroudi, A. (2003) "A new spectral transformation for speaker normalization." *P. EuroSpeech* 1865-1868.

Dusan, S. and Rabiner, L.R. (2005) "Can automatic speech recognition learn more from human speech perception?" *P. 3rd Conf. Speech Tech. Hum.-Comput. Dialogue* 21-36.

Eide E and Gish H (1996) "A parametric approach to vocal tract length normalization." *P. Int. C. Audition, Speech, Signal Process.* 1, 346-348.

Eklund, I. and Traunmüller, H. (1997) "Comparative study of male and female whispered and phonated versions of the long vowels of Swedish." *Phonetica* 54, 1-21.

Fant, G. (1973) "Stops in CV syllables." In *Speech Sounds and Features*, edited by G. Fant. MIT Press, Cambridge, MA, 110-139.

Fazl, A., Grossberg, S. and Mingolla, E. (2008) "View-invariant object category learning, recognition, and search: How spatial and object attention are coordinated using surface-based attentional shrouds." *Cognitive Psychology*, in press.

Ferreira, A.J.S. (2007) " Static features in real-time recognition of isolated vowels at high pitch." *J. Acoust. Soc. Am.* 122(4), 2389-2404.

Fishman, Y.I., Reser, D.H., Arezzo, J.C., and Steinschneider, M. (1998) "Pitch vs. spectral encoding of harmonic complex tones in primary auditory cortex of the awake monkey." *Brain Res.* 786(1-2), 18-30.

Formisano, E., Kim, D.-S., Di Salle, F., van de Moortele, P.-F., Ugurbil, K., and Goebel, R. (2003) "Mirror-symmetric tonotopic maps in human primary auditory cortex" *Neuron* 40(4), 859-869.

Glavitsch, U. (2003) "Speaker normalization with respect to F0: A perceptual approach." *TIK Report No 185*, Swiss Federal Institute of Technology Zurich.

Glasberg, B.R. and Moore, B.C. (1990) "Derivation of auditory filter shapes from notched-noise data." *Hear. Res.* 47(1-2), 103-138.

Goldinger, S.D. (1996) "Words and voices: Episodic traces in spoken word identification and recognition memory." *J Exp. Psychol. Learn. Mem. Cogn.* 22, 1166-1183.

Goldinger, S.D. (1997) "Words and voices: Perception and production in an episodic lexicon." In Johnson K and Mullennix JW (eds) *Talker Variability in Speech Processing*. San Diego: Academic Press, 33-66.

Goldinger, S.D. and Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. J. Phonetics, 31, 305-320.

Goodale, M.A. and Milner, D. (1992) "Separate visual pathways for perception and action." Trends in Neurosci., 15, 10-25.

Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Math.*, 52, 213-257.

Grossberg, S. (1976a) "Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors." *Biol. Cybern.* 23, 121-134.

Grossberg, S. (1976b) "Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, illusions." *Biol. Cybern.* 23, 187-202.

Grossberg, S. (1978) "A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans." In R Rosen and F Snell (Eds) *Progress in theoretical biology, Volume 5***.** New York: Academic, 233-374.

Grossberg, S. (1980) "How does a brain build a cognitive code?" *Psychol. Rev.* 87, 1-51.

Grossberg, S. (1994) "3-D vision and figure ground separation by visual cortex." *Percept. Psychophys.* 55(1), 48-120.

Grossberg, S. (1999) "The link between brain learning, attention, and consciousness." *Conscious. Cogn.* 8, 1-44.

Grossberg, S. (2000) "The complementary brain: Unifying brain dynamics and modularity." *Trends Cogn. Sci.* 4**,** 233-246.

Grossberg, S. (2003a) "How does the cerebral cortex work? Development, learning, attention, and 3D vision by laminar circuits of visual cortex." *Behav. Cogn. Neurosci. Rev.* 2, 47-76.

Grossberg, S. (2003b) "Resonant neural dynamics of speech perception." *J. Phon.* 31, 423-445.

Grossberg, S., Boardman, I., and Cohen, M. (1997) "Neural dynamics of variable-rate speech categorization." *J. Exp. Psychol. Human* 23, 418-503.

Grossberg, S., Govindarajan, K.K., Wyse, L., and Cohen, M.A. (2004) "ARTSTREAM: A neural network model of auditory scene analysis and source segregation." *Neural Networks* 17, 511-536.

Grossberg, S. and Merrill, J.W.L. (1996) "The hippocampus and cerebellum in adaptively timed learning, recognition, and movement." *J. Cogn. Neurosci.* 8, 257-277.

Grossberg, S. and Myers, C.W. (2000) "The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects." *Psychol. Rev.* 107, 735-767.

Grossberg, S. and Repin, D. (2003). "A neural model of how the brain represents and compares multi-digit numbers: spatial and categorical processes". *Neural Networks*, 16, 1107-1140.

Grossberg, S. and Stone, G.O. (1986) "Neural dynamics of attention switching and temporal order information in short-term memory." *Mem. Cognition* 14, 451-468.

Grossberg, S. and Versace, M. (2008) "Spikes, synchrony, and attentive learning by laminar thalamocortical circuits." *Brain Research* (in press).

Grossberg, S. and Williamson, J.R. (1999) "A self-organizing neural system for learning to recognize textured scenes." *Vision Res.* 39, 1385-1406.

Grunewald, A. and Grossberg, S. (1998) “Self-organization of binocular disparity tuning by reciprocal corticogeniculate interactions.” *J. Cogn. Neurosci.* 10, 199-215.

Guenther, F.H. (1995) “Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production.” *Psychol. Rev*. 102, 594-621.

Guenther, F.H., Ghosh, S.S. and Tourville, J.A. (2006) “Neural modeling and imaging of the cortical interactions underlying syllable production.” *Brain and Language*, 96, 280-301.

Gutschalk, A., Patterson, R.D., Scherg, M., Uppenkamp, S., and Rupp, A. (2004) “Temporal dynamics of pitch in human auditory cortex.” *NeuroImage* 22, 755-766.

Hackett, T.A., Stepniewska, I., and Kaas, J.H. (1998) “Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys.” *J. Comp. Neur.* 394(4), 475-495.

Hawkins, S. (2003). “Roles and representations of systematic fine phonetic detail in speech understanding.” *J. Phonetics*, 31, 373-405.

Heil, P., Rajan, J., and Irvine, D.R. (1994) “Topographic representation of tone intensity along the isofrequency axis of cat primary auditory cortex.” *Hear. Res.* 76(1-2), 188-202.

Hess, W. (1983) *Pitch Determination of Speech Signals-Algorithms and Devices*. Berlin: Springer.

Hickok, G. and Poeppel, D. (2007) “The cortical organization of speech processing” *Nat. Rev. Neurosci*, 8, 393-402.

Hillenbrand, J.M. and Gayvert, R.T. (1993) “Identification of steady-state vowels synthesized from the Peterson and Barney measurements.” *J. Acoust. Soc. Am.* 94, 668-674.

Hillenbrand, J.M. and Nearey, T.M. (1999) “Identification of resynthesized /hVd/ utterances: effects of formant contour.” *J. Acoust. Soc. Am.* 105(6), 3509-3523.

Hillenbrand, J.M., Houde, R.A., and Gayvert, R.T. (2006) “Speech perception based on spectral peaks versus spectral shape.” *J. Acoust. Soc. Am.* 119(6), 4041-4054.

Holdsworth, J., Nimmo-Smith, I., Patterson, R., and Rice, P. (1988) “Implementing a gammatone filterbank.” *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank.*

Hubel, D. H., and Wiesel, T. N. (1962). “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex.” *J. Physiol*., 160, 106-154.

Hudspeth, A.J. (2000) “Chapter 30: Hearing” *Principles of Neuroscience* 4_{th} *Edition*, eds Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell. New York: McGraw Hill, 590-613.

Imig, T.J., Ruggero, M.A., Kitzes, L.M., Javel, E., and Brugge, J.F. (1977) "Organization of auditory cortex in the owl monkey." *J. Comp. Neur.* 171(1), 111-128.

Ito, M., Tamura, H., Fujita, I. and Tanaka, K. (1995) "Size and position invariance of neuronal responses in monkey inferotemporal cortex." J. Neurophysiol., 73, 218-226.

Johnson, K. (1990) "The role of perceived speaker identity in F0 normalization of vowels." *J. Acoust. Soc. Am.* 88, 642-654.

Johnson, K. (1997a) "Speech perception without speaker normalization: an exemplar model." In Johnson K and Mullennix JW (eds) *Talker Variability in Speech Processing*. San Deigo: Academic Press, 145-166.

Johnson, K. (1997b) "The auditory/perceptual basis for speech segmentation." *OSU Working Papers in Linguistics* 50, 101-113, Columbus, Ohio.

Johnson, K. (2005) "Speaker normalization in speech perception." In Pisoni DB and Remez R (eds) *The Handbook of Speech Perception*. Oxford: Blackwell Publishers, 363-389.

Johnson, K. (2006) "Resonance in an exemplar-based lexicon: The emergence of social identity and phonology." *J.Phon.* 34, 485-499.

Johnson, K., Strand, E.A., and D'Imperio, M. (1999) "Auditory-visual integration of talker gender in vowel perception." *J. Phon.* 27, 359-384.

Kaas, J.H. and Hackett, T.A. (1998) "Subdivisions of auditory cortex and levels of processing in primates." *Audiol. Neuro-Otol.* 3(2-3), 73-85.

Kaas, J.H. and Hackett, T.A. (2000) "Subdivisions of auditory cortex and processing streams in primates." *P. Natl. Acad. Sci.* 97(22), 11793-11799.

Kastner, S. and Ungerleider, L. G. (2001). "The neural basis of biased competition in human visual cortex." *Neuropsychologia,* 39, 1263-1276.

Kato, K. and Kakehi, K. (1988) "Listener adaptability to individual speaker differences in monosyllabic speech perception." *J. Acoust. Soc. Jpn.* 44, 180-186.

Kent, R.D. and Read, C. (1992) *Acoustic Analysis of Speech*. Madison: Singular.

Kirkman, T.W. (1996) *Statistics to Use*. http://www.physics.csbsju.edu/stats (1 Oct 2007).

Kraljic, T. and Samuel, A.G. (2007) "Perceptual adjustments to multiple speakers." *J. Mem. Lang.* 56, 1-15.

Langner, G., Sams, M., Heil, P., and Schulze, H. (1997) "Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: Evidence from magnetoencephalography." *J. Comput. Physiol.* 181(6), 665-676.

Lee, L. and Rose, R. (1996) "Speaker normalization using efficient frequency warping procedures." *P. Int. C. Audition, Speech, Signal Process.* 1, 353-356.

Lee, L. and Rose, R. (1998) "A frequency warping approach to speaker normalization." *IEEE Trans. Speech Audio Process.* 6(1), 49-60.

Lehiste, I. and Meltzer, D. (1973) "Vowel and speaker identification in natural and synthetic Speech." *Lang. Speech* 16, 356-264.

Lindau, M. (1978) "Vowel features." *Language* 54(3), 541-563.

Lloyd, R.J. (1890a) *Some Researches into the Nature of Vowel-Sound*, Liverpool, England: Turner and Dunnett.

Lloyd, R.J. (1890b) "Speech sounds: Their nature and causation (I)." *Phonetische Studien* 3, 251-278.

Lloyd, R.J. (1891) "Speech sounds: Their nature and causation (II-IV)." *Phonetische Studien* 4, 37-67, 183-214, 275-306.

Lloyd, R.J. (1892) "Speech sounds: Their nature and causation (V-VII)." *Phonetische Studien* 5, 1-32, 129-141, 263-271.

Lockwood, A.H., Salvi, R.J., Coad, M.L., Arnold, S.A., Wack, D.S., Murphy, B.W., and Burkard, R.F. (1999) "The functional anatomy of the normal human auditory system: Responses to 0.5 and 4.0 kHz tones and varied intensities." *Cereb. Cortex* 9(1), 65-76.

Luethke, L.E., Krubitzer, L.A., and Kaas, J.H. (1988) "Cortical connections of electrohpysiologically and architectonically defined subdivisions of auditory cortex in squirrels." *J. Comput. Neur.* 268(2), 181-203.

Magimai-Doss, M., Stephenson, T.A., Bourlard, H. (2003) "Using pitch frequency information in speech recognition." In *P. Eurospeech* 2003.

McDonough, J. and Byrne, W. (1999) "Speaker adaptation with all-pass transforms." *P. Int. C. Audition, Speech, Signal Process.* 2, 757-760.

Merzenich, M.M. and Brugge, J.F. (1973) "Representation of the cochlear partition of the superior temporal plane of the macaque monkey." *Brain Res.* 50, 275-296.

Miller, J. (1989) "Auditory-perceptual representation of the vowel." *J. Acoust. Soc. Am.* 85, 2114-2134.

Moore, C.B. and Jongman, A. (1997) "Speaker normalization in the perception of Mandarin Chinese tones." *J. Acoust. Soc. Am.* 102, 1864-1877.

Morel, A. and Kaas, J.H. (1992) "Subdivisions and connections of auditory cortex in owl monkeys." *J. Comput. Neur.* 318, 27-63.

Morel, A., Garraghty, P.E., and Kaas, J.H. (1993) "Tonotopic organization, architectonic fields, and connections of auditory cortex in macaque monkeys." *J. Comput. Neur.* 335, 437-459.

Nearey, T.M. (1989) "Static, dynamic, and relational properties in vowel perception." *J. Acoust. Soc. Am.* 85, 2088-2113.

Nearey, T.M., Hogan, J., and Rozsypal, A. (1979) "Speech signals, cues and features." In *Perspectives in Experimental Linguistics*, edited by G. Prideaux, Benjamin, Amsterdam.

Piaget, J. (1963). *The Origins of Intelligence in Children.* New York: Norton.

Page, M. (2000). "Connectionist modellino in psychology: A localist manifesto." *Behav. and Brain Sci.*, 23, 443-467.

Palmeri, T.J., Goldinger, S.D., and Pisoni, D.B. (1993) "Episodic encoding of voice attributes and recognition memory for spoken words." *J. Exp. Psychol. Learn. Mem. Cogn.* 19(2): 309-328.

Pantev, C., Hoke, M., Lehnertz, K., Lutkenhoner, B., Anogianakis, G., and Wittkowski, W. (1988) "Tonotopic organization of the human auditory cortex revealed by transient auditory evoked magnetic fields." *Electroencephalography Clin. Neurophysiol.* 69(2), 160-170.

Pantev, C., Hoke, M., Lutkenhoner, B., and Lehnertz, K. (1989) "Tonotopic organization of the auditory cortex: pitch versus frequency representation." *Science* 246(4929), 486-488.

Patterson, R., Nimmo-Smith, I., Holdsworth, J., Rice, P. (1987) "An efficient auditory filterbank based on the gammatone function." *Annex B of the SVOS Final Report: Part A: The Auditory Filterbank.*

Patterson, R., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1988) *Spiral VOS Final Report: Part A: The Auditory Filterbank.*

Patterson, R.D. and Rice, P. (1987) "A preliminary study of the feasibility of a hardware version of the auditory filterbank" *Annex A of the SVOS Final Report: Part A: The Auditory Filterbank.*

Patterson, R.D., Uppenkamp, S., Johnsrude, I.S., and Griffiths, T.D. (2002) "The processing of temporal pitch and melody information in auditory cortex." *Neuron* 36(4), 767-776.

Penagos, H., Melcher, J.R., and Oxenham, A. (2004) "A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging." *J.*

Neurosci. 24(30), 6810-6815.

Peterson, G.E. (1961) “Parameters of vowel quality.” *J. Speech Hear. Res.* 4, 10-29.

Peterson, G.E. and Barney, H.L. (1952) “Control methods used in a study of the vowels.” *J. Acoust. Soc. Am.* 24, 175-184.

Petkov, C.I., Kayser, C., Augath, M., and Logothetis, N.K. (2006) “Functional imaging reveals numerous fields in the monkey auditory cortex.”*PLOS*4 e215.

Pierrehumbert, J. (2006) “The next toolkit.” *J. Phon.* 34, 516-530.

Plack, C.J., Oxenham, A.J., Popper, A.N., and Fay, R.R. (2005) *Pitch: Neural Coding and Perception*. New York: Springer Verlag.

Qin, L., Sakai, M., Chimoto, S. and Sato, Y. (2005) “Interaction of excitatory and inhibitory frequency-receptive fields in determining fundamental frequency sensitivity of primary auditory cortex neurons in awake cats.” *Cereb. Cortex* 15, 1371-1383.

Rabiner, L.R. and Shafer, R.W. (1978) *Digital Processing of Speech Signals* Edgewood Cliffs, New Jersey: Prentice-Hall.

Ragot, R. and Lepaul-Ercole, E. (1996) “Brain potentials as objective indexes of auditory pitch extraction from harmonics.” *Neuroreport* 7(4), 905-909.

Rauschecker, J.P. and Tian, B. (2004) “Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey.” *J. Neurophysiol.* 91, 2578-2589.

Rauschecker, J.P., Tian, B., and Hauser, M. (1995) “Processing of complex sounds in the macaque nonprimary auditory cortex.” *Science* 268(5207), 111-114.

Reale, R.A. and Imig, T.J. (1980) “Tonotopic organization in the auditory cortex of the cat.” *J. Comput. Neur.* 192(2), 265-291.

Romani, G.L., Williamson, S.J., and Kaufman, L. (1982) “Tonotopic organization of the human auditory cortex.” *Science* 216(4552), 1339-1340.

Schwippert, C. and Benoit, C. (1997) “Audiovisual intelligibility of an androgynous speaker.” In *P. ESCA Workshop Audio Vis. Speech Process. (AVSP ’07): Cogn. Comput. Approaches*, Rhodes, Greece (Benoit C and Campbell R, editors), 81-84.

Schulze, H., Hess, A., Ohl, F., and Scheich, H. (2002) “Superposition of horseshoe-like periodicity and linear tonotopic maps in auditory cortex of the Mongolian gerbil.” *Eur. J. Neurosci.* 15(6), 1077-1084.

Seldon, H.L. (1985) “The anatomy of speech perception: Human auditory cortex.” In Peters, A and Jones EG (eds) *Cerebral Cortex* 4, New York: Plenum Press, 273-327.

Slaney, M. (1993) “An efficient implementation of Patterson-Holdsworth auditory filter bank.” *Apple Computer Technical Report*, #35.

Slaney, M. (1998) “Auditory toolbox, version 2.” *Interval Research Corporation Technical Report* #10.

Slawson, A.W. (1968) “Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency.” *J. Acoust. Soc. Am.* 43, 87-101.

Spitzer, H., Desimone, R., and Moran, J. (1998) “Increased attention enhances both behavioral and neuronal performance.” *Science* 240, 338-340.

Stevens, K.N. (1998) *Acoustic Phonetics*. Cambridge: MIT Press.

Strand, E.A. and Johnson, K. (1996) “Gradient and visual speaker normalization in the perception of fricatives.” In Gibbon D. (ed) *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielefeld*. Berlin: Mouton de Gruyter, 14-26.

Summerfield, Q. and Haggard, M.P. (1973) “Vocal tract normalization as demonstrated by reaction times.” *R. Speech Res. Progress* 2, 1-12, The Queen’s University of Belfast, Belfast, Ireland.

Sussman, H.M. (1986) “A neuronal model of vowel normalization and representation.” *Brain Lang.* 28(1), 12-23.

Sussman, H.M., Bessell, N., Dalston, E., and Majors, T. (1997) “An investigation of stop place of articulation as a function of syllable position.” *J. Acoust. Soc. Am.* 101(5), 2826-2838.

Syrdal, A.K. and Gopal, H.S. (1986) “A perceptual model of vowel recognition based on the auditory representation of American English vowels.” *J. Acoust. Soc. Am.* 79, 1086-1100.

Talavage, T.M., Ledden, P.J., Benson, R.R., Rosen, B.R., and Melcher, J.R. (2000) “Frequency-dependent responses exhibited by multiple regions in human auditory cortex” *Hear. Res.* 150, 225-244.

Talavage, T.M., Sereno, M.I., Melcher, J.R., Ledden, P.J., Rosen, B.R., and Dale, A.M. (2004) “Tonotopic organization in human auditory cortex reveled by progressions of frequency sensitivity” *J. Neurophysiol.* 91, 1282-1296.

Titze, I.R. (1994) “Mechanical stress in phonation.” *J. Voice* 8(2), 99-105.

Traunmüller, H. (1981) “Perceptual dimension of openness in vowels.” *J. Acoust. Soc. Am.* 69, 1465-1475.

Tunturi, A.R. (1952) “A difference in the representation of auditory signals from the left and the right ears in the isofrequency of the right middle ectosylvian auditory cortex of the dog.” *Am. J. Physiol.* 168, 712-727.

Turner, R.E. and Patterson, R.D. (2003) “An analysis of the size information in classical formant data: Peterson and Barney (1952) revisited.” J. *Acoust. Soc. Jpn.* 33(9), 585-589.

Ungerleider, L.G. and Mishkin, M. (1982) “Two cortical visual systems: Separation of appearance and location of objects. In Ingle DL, Goodale MA, and Mansfield RJW (eds) Cambridge MA: MIT Press, 549-586.

Verbrugge, R.R., Strange, W., Shankweiler, D.P., and Edman, T.R. (1976) “What information enables a listener to map a talker’s vowel space?” *J. Acoust. Soc. Am.* 60, 198-212.

Vitevitch, M.S. and Luce, P.A. (1991). Probabilistic phonotactics and neighborhood activation in spoken word recognition. J. Memory and Lang., 40, 374-408.

Walker, S., Bruce, V., and O’Malley, C. (1995) “Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect.” *Percept. Psychophys.* 57, 1124-1133.

Watrous, R.L. (1991) “Current Status of Peterson-Barney vowel formant data.” *J. Acoust. Soc. Am.* 89, 2459-2460.

Wegman, S., McAllaster, D., Orloff, J., and Peskins, B. (1996) “Speaker normalization on conversational telephone speech.” *P. Int. C. Audition, Speech, Signal Process.* 1, 339-341.

Wessa, P. (2007). *Free Statistics Software*, Office for Research Development and Education. Version 1.1.22-rl. http://www.wessa.net (1 Oct 2007).

Wessinger, C.M., Buonocore, M.H., Kussmaul, C.L., and Mangun, G.R. (1998) “Tonotopy in human auditory cortex examined with functional magnetic resonance imaging.” *Hum. Brain Mapp.* 5(1), 18-25.

Whitfield, I.C. (1980) “Auditory cortex and the pitch of complex tones.” *J. Acoust. Soc. Am.* 67(2), 644-647.

Yazdanbakhsh, A. and Grossberg, S. (2004) “Fast synchronization of perceptual grouping in laminar visual cortical circuits.” *Neural Networks* 17, 707-718.

Zahorian, S.A. and Jagharghi, A.J. (1991) “Speaker normalization of static and dynamic vowel spectral features.” *J. Acoust. Soc. Am.* 90(1), 67-75.

Zhan, P. and Westphal, M. (1997) “Speaker normalization based on frequency warping.” *P. Int. C. Audition, Speech, Signal Process.* 2, 1039-1041.

Zhan, P. and Waibel, A. (1997) “Vocal tract length normalization for large vocabulary continuous speech recognition.” *Technical Report CMU-CS-97-148*, School of Computer Science, Carnegie Mellon University.

Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J.J. (2007) “Trade-off between object selectivity and tolerance in monkey inferotemporal cortex.” J. Neurosci. 26, 13025-13026.