First publ. in: Journal of the Acoustical Society of America, 119 (2006), 6, pp. 4006-4015

Evidence for attractors in English intonation

Bettina Braun,[a] Greg Kochanski, Esther Grabe, and Burton S. Rosner
Phonetics Laboratory, University of Oxford, 41 Wellington Square, Oxford OX1 2JF, United Kingdom

Although the pitch of the human voice is continuously variable, some linguists contend that intonation in speech is restricted to a small, limited set of patterns. This claim is tested by asking subjects to mimic a block of 100 randomly generated intonation contours and then to imitate themselves in several successive sessions. The produced f_0 contours gradually converge towards a limited set of distinct, previously recognized basic English intonation patterns. These patterns are "attractors" in the space of possible intonation English contours. The convergence does not occur immediately. Seven of the ten participants show continued convergence toward their attractors after the first iteration. Subjects retain and use information beyond phonological contrasts, suggesting that intonational phonology is not a complete description of their mental representation of intonation.

PACS number(s): 43.70.Fq, 43.71.Bp, 43.66.Ba, 43.66.Fe [AL] Pages: 4006–4015

I. INTRODUCTION

The pitch of the human voice is continuously variable. Nevertheless, phonologists often assert that any language uses only a small set of different patterns to control intonation (variation in pitch, whose primary acoustic correlate is fundamental frequency or f_0). Intonation in English, for example, is said to behave this way (Cruttenden, 1997; Kingdon, 1958; O'Connor and Arnold, 1961). Similar claims have been made for numerous other languages. See Hirst and Cristo (1998) for discussions of intonation patterns in European and non-European languages.

Listeners supposedly interpret and make linguistic sense of continuous pitch changes in speech via such basic patterns (Brazil, 1985; Jun, 2005; Pierrehumbert and Hirschberg, 1990). In intonation languages, such as Dutch, German, or English, pitch variations seem to help in signaling prosodic phrasing, different information structure packaging (Steedman, 2000; Vallduví and Engdahl, 1996), or attitudinal and emotional information (Scherer, 1985). The belief in a small set of basic intonation patterns is based primarily on a linguist's or a subject's conscious classification of contours (Gussenhoven and Rietveld, 1997; Kohler, 1991; Ladd and Morton, 1997). Experimental evidence on the validity of linguistic descriptions of intonation is very limited.

One line of attempts to obtain such evidence has used the "categorical perception" paradigm (Liberman *et al.*, 1957). A set of speech stimuli is generated whose f_0 contours are spaced along a continuum between two supposedly basic prototypes. In an identification task, participants classify each stimulus as belonging to one of the prototypes. In a discrimination task, the observers indicate whether they hear a difference between paired stimuli that are near neighbors on the continuum. The hallmark of "categorical perception" is that discrimination can be predicted from identification. Qualitatively, all examples within a category should be perceived as similar, making it difficult to discriminate pairwise between them. In contrast, discrimination should be easy for two items that straddle the boundary between categories, producing peak performance.

A maximum in the discrimination function was reported for early and medial peaks in German intonation (Kohler, 1991) and for the perception of high and low boundary tones in Dutch (Remijsen and van Heuven, 1999). However, discrimination within categories was better than the identification data predicted. In another experiment, Ladd and Morton (1997) tested the perception of peak height. Their listeners could classify the stimuli as normal or unusual, but discrimination was not even maximal across the classification boundary. In studies on lexical tone in Cantonese, the predictability of discrimination from identification varied with the type of contrast under study (Francis *et al.*, 2003). On the whole, these data show no clear examples of categorical perception.

Obviously, identification relies on conscious classification of the speech stimuli. Furthermore, it shows nothing about whether the ends of the continuum represent basic psychological structures or are simply transient categories imposed by the experiment. Color naming provides a case in point. In a paper by Doll and Thomas (1967), subjects were trained to label two different wavelengths and were then tested on intermediate wavelengths, to generate an identification function. Training on different pairs of wavelengths then resulted in different identification functions. Similar effects occur in speech perception (cf. Eisner and McQueen, 2005 and references therein). Such easily shifted category boundaries apparently cannot be deeply embedded in our perceptions. Indeed, boundaries can move so rapidly (Ladefoged, 1989; Ladefoged and Broadbent, 1957) that they might not even be stable over the duration of a categorical perception experiment. For these and other reasons, the concept of categorical perception and its attendant experimental paradigm have become increasingly disfavored [see Schouten *et al.* (2003) and Plomp (2002, pp. 137–145)].

In a different approach, Pierrehumbert and Steele (1989) constructed a set of contours that varied in the position of a

[a]Present address: Max Planck Institute for Psycholinguistics, Postbus 310 NL-6500 AH Nijmegen.

peak in f_0. They asked subjects to repeat each carrier utterance and measured the positions of the resulting peaks. Although the paper lacks a statistical analysis, the authors interpret the data as showing that the subjects produced bimodal distributions of peak positions. The paper introduced a valuable method, although it was limited in certain respects. Beyond the lack of statistics, one of the five subjects was an author and was therefore not naïve.

These previous experimental efforts have left undefined the number and properties of basic intonational patterns. Even their psychological reality can be doubted. We therefore set out to obtain better behavioral evidence on whether basic intonational patterns affect the perception and production of speech. To do this, we employ iterative mimicry. Mimicry is a simple behavioral response to language. It appears early in children's language development, long before grammar and comprehension are fully established (Loeb and Allen, 1993; Meltzoff and Decety, 2003; Snow, 1998). Results from mimicry therefore should provide a better picture of a subject's language processing than do conclusions based on introspection or explicit and conscious classification.

Bartlett (1932) long ago exploited a version of iterative mimicry to study drawing, among other things. His Method of Serial Reproduction required a subject to produce a drawing after short exposure to an original. The reproduction was presented to another subject who attempted a new reproduction. This between-subjects procedure was iterated some dozen or so more times. Bartlett found that the successive reproductions were gradually simplified or even transformed into something very different from the original drawing.

Our procedure of iterative mimicry builds on this method but remains strictly within each subject. It provides a substantially new approach to studying the way in which intonation is processed. Our technique is related to those used by Pierrehumbert and Steele (1989) and Repp and Williams (1987), but it extends their efforts in several ways. In our procedure, a subject first mimics an initial set of English utterances with widely varied f_0 contours. In a second session, the subject then mimics his/her own first productions. Continuing in two more sessions, he/she mimics his/her own productions from the immediately preceding session. If basic intonational contours are part of the subject's mental structures for English, the contours in the subject's productions should progressively reduce over sessions towards a few, well-distinguished forms.

Such well-distinguished contours can be idealized mathematically as "attractors," a concept from the mathematics of iterated functions. A function $f(x)$ is applied to a starting value x_0 and then successively to its own result, yielding a sequence of values x_0, $x_1=f(x_0)$, $x_2=f(x_1)$, …. A simple example of an attractor comes from the function $f(x)=x^2$. Zero is called a fixed point of this function, since $f(0)=0$. [Generally, w is a fixed point of a function g if $g(w)=w$.]

Consider values of the argument of $f(x)$ in the region around zero, where $|x_0|<1$. Applying $f(x)$ to any value in that region yields a result that has a smaller absolute value than the input value. Succeeding applications of the function produce still smaller values that approach zero as the sequence continues. For the function x^2, then, zero is a *stable* fixed point. Nearby values will converge over iterations toward any stable fixed point. This convergence is a key feature of an attractor. The points from which the sequence converges form an attractor's "basin of attraction."

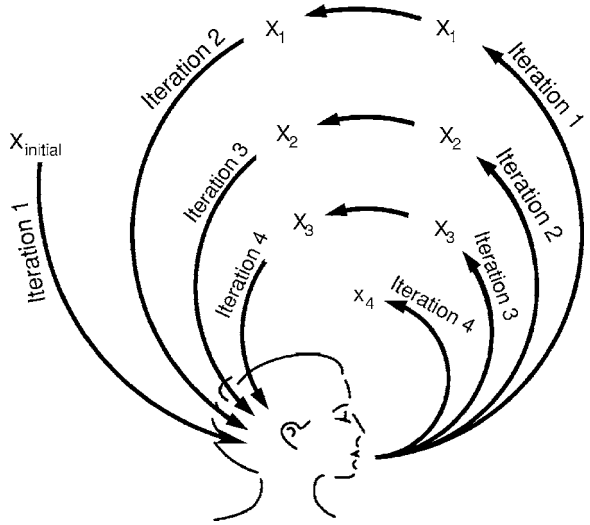

FIG. 1. Scheme of the experiment. The numbers refer to the four experimental sessions, called iterations. In iteration 1, the subject mimics the initial synthesized stimuli X_{initial}, producing responses x_1. In each later iteration, the subject mimics his or her own productions from the preceding session. Each response serves as a stimulus in the next iteration.

For iterative mimicry, x represents a description of an entire intonation contour as a vector of f_0 values for each moment in time. Then $f(x)$ represents the transformation from the sound entering the subject's ears to the sound produced by the subject. If each x in a subset of starting contours converges over successive imitations towards a fixed shape, this would provide evidence for the existence of an attractor in intonation. That subset of contours would lie in a basin of attraction. Different subsets of starting contours that converged towards different fixed shapes would lie in different basins of attractions. The attractors themselves would represent underlying mental structures.

Using iterative mimicry to search for underlying attractors has a crucial advantage over the other experimental methods described thus far: it does not require any conscious classification of intonation by either the subject or the experimenter. This experimental procedure comes much closer than others to the actual use of language in conversation. We respond to language, but rarely do we consciously and explicitly analyze the intonation we hear. Hence, our method avoids the possibility that conscious reports on intonation may not correspond to responses that speakers would make in a conversation.

II. PROCEDURES

A. General design

We designed the experiment in analogy with the iterated-function definition of attractors given above. Figure 1 shows this design. Each subject serves in four experimental sessions. In the first session, "iteration 1," the subject mimics an initial set of utterances X_{initial}. The f_0 contours of these utterances are systematically varied by resynthesis. Male and

female subjects hear initial stimuli that are based on a male and a female voice, respectively. We record the subject's response to each X_{initial}. We call this first set of responses x_1. These responses become the stimuli X_1 for the next session.

In that session, the subject mimics each utterance in X_1, producing a second set of responses x_2. These responses in turn become the stimuli X_2 for the next session. The responses x_3 from that session become the stimuli X_3 for the final session, which is "iteration 4." In that iteration, the subject mimics each utterance in X_3, yielding the final set of responses x_4. Except for the initial stimuli X_{initial}, a subject hears only his or her own utterances from the preceding iteration.

During each iteration, the subject needs to remember the intonation pattern for only a short time between stimulus and mimicry. A stimulus is presented to the subject who begins to respond immediately when ready. The mean interval within a session between the end of a stimulus and the beginning of the response to it was 130 ms with a standard deviation of 55 ms. Subjects therefore typically started moving their articulators before a stimulus ended. The median spacing between iterations (i.e., x_k to X_k) was 5 days, with a minimum spacing of 1 day. Additionally, we scrambled the order of presentation of stimuli between iterations so that subjects would not be able to track an utterance from one iteration to the next.

B. Participants

Participants were linguistically naïve speakers of Standard Southern British English, five males and five females. All had normal hearing and were between 19 and 30 years old. They received written instruction to imitate the speech and the melody of each sentence as closely as possible. They were informed that the initial stimuli were synthesized but that their task was not to imitate the voice *quality* of the stimuli. To ease any discomfort at hearing their own voice in the self-mimicking sessions, subjects were told that they would hear processed versions of their own speech. (See Sec. II E for details of the processing.)

C. Materials

The target sentences were chosen to be mostly sonorant subject-verb-object constructions with a total of seven syllables each. They were constructed so that in a neutral reading, accents occur on the first and the sixth syllables. Nine candidate sentences were generated.

D. Initial recording and selection

We interspersed these nine candidates with fillers of various lengths and syntactic structures, in order to obtain starting utterances for resynthesis. One male and one female speaker of Southern British English read the list, both with a rising (e.g., questioning) and falling (i.e., statement) final intonation. The recordings were made in a sound-treated room and digitized at 16 kHz with 16 bits/sample. The productions from the male and the female speaker were processed separately to provide initial stimuli for the male and female subjects, respectively. Having two versions of the recordings minimized the eventual size of the f_0 shifts that we would later need to impose; resynthesis quality generally declines as f_0 shifts become larger.

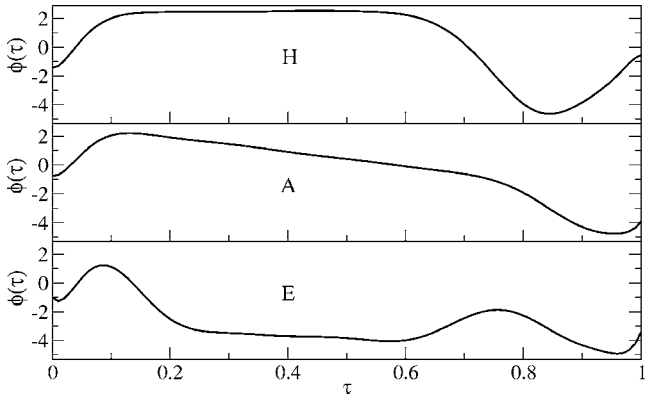

FIG. 2. The three basis contours from which initial stimuli were constructed (Grabe *et al.*, 2003). The ordinate is in semitones, relative to mean f_0; the abscissa is normalized time.

The nine candidate sentences were PSOLA-resynthesized with the three basis intonation contours A, E, and H of Grabe *et al.* (2003), see Fig. 2. (Resynthesis of A and E was based on the falling recordings, H on the rising recordings.) The basis contours were chosen to be maximally distinct among normal English contours. The authors separately assessed the quality and naturalness of the 54 resynthesized versions. Six sentences survived this initial selection.

Next, to avoid any possible bias toward either rising or falling patterns (Ryalls *et al.*, 1994), we constructed versions of these six sentences that were intermediate between the rising and falling recordings in terms of duration and LPC coefficients (see the Appendix). One more sentence was eliminated because the resulting audio had an unnatural segmental structure. This finally left the following five sentences for the main experiment, each with a male and female version:

(1) Anna will marry Marlon.
(2) Alan rode on a llama.
(3) Eleanor blamed our neighbour.
(4) Melanie won a million.
(5) Minnie will ring a lawyer.

1. Intonation contours for initial stimuli: $X_{(\text{initial},k)}$

One hundred fifteen initial f_0 contours were generated by taking linear combinations of the basis contours shown in Fig. 2. A combination was defined by $c_A \cdot a(\tau) + c_E \cdot e(\tau) + c_H \cdot h(\tau)$, where $a(\tau)$, $e(\tau)$, and $h(\tau)$ are the basis contours expressed as functions of normalized time τ that ranges from 0 to 1. We used combinations of c_A, c_E, and c_H such that $c_A + c_E + c_H = 1$, $c_A > -0.3$, $c_E > -0.3$, and $c_H > -0.3$.

The 115 combinations of c_A, c_E, and c_H were selected to avoid clusters. The selection algorithm operated iteratively. At each step, it chose 100 candidate samples of (c_A, c_E, c_H) from a uniform distribution and accepted the candidate that was farthest from all previously accepted samples. (Farthest is defined via the Euclidean distance between the two

(c_A, c_E, c_H) vectors: $[(C_A - c'_A)^2 + (c_E - c'_E)^2 + (c_H - c'_H)^2]^{1/2}$.) This resulted in a set of points that were more uniformly distributed than would result from independent random sampling. To avoid priming the subjects with the basis contours, we excluded regions where (c_A, c_E, c_H) was within 0.1 of (1,0,0), (0,1,0), or (0,0,1). This process finally produced initial intonation contours that were generally intermediate between the basis contours (when $c_A > 0$, $c_E > 0$, and $c_H > 0$) but also included some mildly exaggerated versions of the basis contours when $c_A < 0$, $c_E < 0$, or $c_H < 0$.

2. Synthesizing the initial stimuli

Each initial intonation contour was superimposed on a segmentally intermediate target sentence from the male and the female speaker, using PSOLA resynthesis. Although each subject heard the resynthesized male or female initial utterances with the same initial contours, f_0 was scaled up or down from those utterances to match the average f_0 of the participant's own speech. For the ten subjects, this ultimately produced 1150 synthetic utterances.

E. Data collection

Each experimental session used up to five successive blocks of stimuli:

- Block A: A practice block of 15 stimuli (generated per Sec. II D 1). These synthetic stimuli were the same on all iterations. Each experimental sentence was synthesized with three distinct contours.
- Block B: Re-recordings of unusable productions detected after the end of the previous iteration. We re-recorded when the subject spoke the wrong words or substantially hesitated in the midst of a sentence, when bursts of noise interfered (e.g., the subject touched the microphone), or when occasional technical problems arose.
- Block C: A block to test reproducibility. In the first iteration, this was identical to block A. In succeeding iterations, the stimuli in this block were always the productions of block C, iteration 1. The productions were adjusted to a consistent amplitude, and initial or final breath and lip-smack noises were removed before they were used as stimuli.
- Block D: This was the main experimental block. The first iteration utilized 100 synthesized stimuli. Each of the five finally selected sentences was resynthesized with 20 maximally separated contours to cover the space of possible contours. Each sentence carried a different block of contours. In all suceeding iterations, the stimuli were the productions of block D from the previous iteration. The productions were adjusted to a consistent amplitude, and initial or final breath and lip-smack noises were removed before they were used as stimuli. We randomized the order of the stimuli between each iteration so that the subjects could not keep track of the history of each stimulus.
- Block E: This block re-recorded blatant mistakes that the experimenter noticed while responses to block D were being recorded.

If a subject was dissatisfied with a production, one repeat was immediately available. Eleven percent of the utterances were re-recorded at once for this reason. Between blocks B and E, an additional 5% of the productions were re-recorded.

With 115 stimuli per session, 4 sessions, and 10 participants, the corpus contains 5200 responses in total. Of these, the 4000 utterances from blocks D (with some replacements from by B and E) are the basis of the results presented below.

F. Signal processing for mimicry data

Signal processing of the mimicry productions involved three main steps: inspection and modification of f_0 tracks, weighting, and normalization. The processing generally followed Kochanski *et al.* (2005). Before analysis, the f_0 tracks obtained from `get_f0` (Talkin, 1995) were screened for gross errors. Tracks that had frequency shifts of one octave, (±15 Hz) between successive points were inspected, as were tracks with points that fell more than 7 semitones below the participant's average f_0. Of the 4600 utterances from blocks C and D, 125 tracks required inspection. Where necessary, f_0 was modified. In 87 utterances some region was raised by an octave. It was lowered by an octave in one. One hundred seventeen productions had the marking of some region changed from voiced to unvoiced or irregularly voiced. The marked regions contributed to an indicator of irregular voicing $I(t)$, used in Eq. (1) below. The mean length of all modified or marked regions was 35 ms, totaling less than 1% of the data.

1. Weighting the mimicry data

For most sonorant sounds, especially in stressed syllables, the perceived pitch of speech correlates well with the output of algorithms that estimate fundamental frequency. Not all our data, however, meet that criterion. In the interest of using plausible sentences, we chose some that had incomplete sonority (e.g., after /d/ in "rode," sentence 2, above). Furthermore, some voiced sounds were not strongly periodic and may not have had a clear pitch.

Consequently, not all the data are of equal value for specifying the pitch that the subject heard and attempted to reproduce. A weight function is unavoidable, because pauses have no pitch and must therefore be excluded from further analysis.

We base the weight, $W(t)$, on two local acoustic measures that seem important in describing prosody: loudness, $L(t)$, and the degree of periodicity of the waveform, $A(t)$. The net result of $W(t)$ is to focus the analyses on the peak of the syllable, paying less attention to the margins. The choice of weight is partially motivated by the pitch tracking algorithm and statistical considerations. Another factor is that low-amplitude parts of speech are often buried in normal environmental noise and thus have little or no perceptual importance. [Substantial numbers of people speak in environments where the mean signal-to-noise ratio for speech is only 9 dB (Kochanski *et al.*, 2005).]

The weight of a datum is

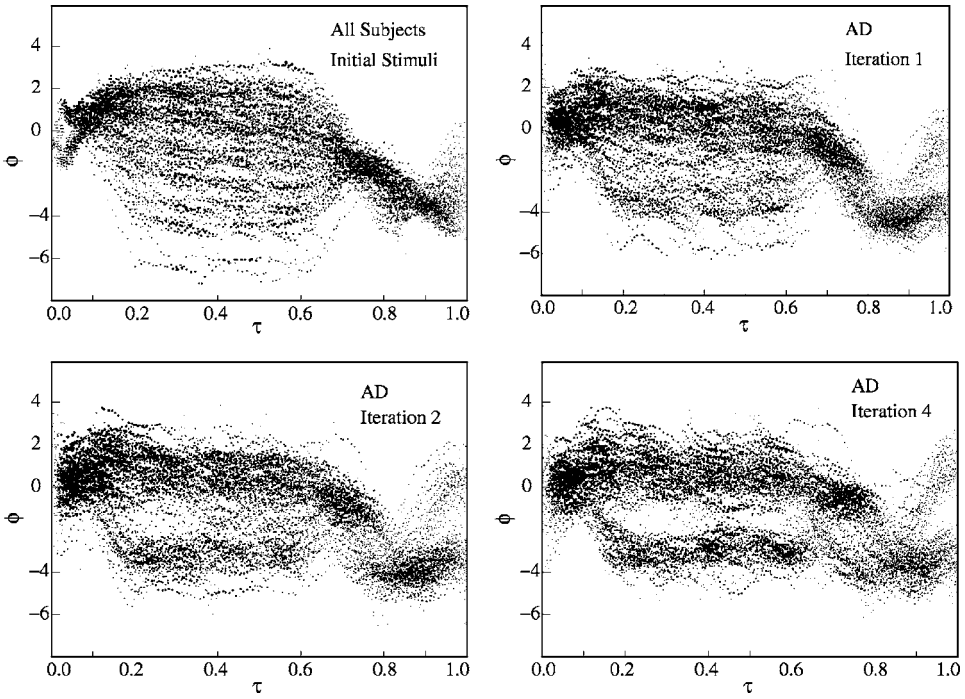

FIG. 3. Normalized f_0 contours, $\phi(\tau)$, for a typical subject, AD. In each panel, f_0 measurements are superposed from the 100 main experimental utterances. The time axis τ is normalized to the length of the utterance and the frequency axis is semitones relative to the speaker's mean f_0. The upper left panel shows f_0 contours for the initial stimuli. The next three panels contain response tracks for iterations 1 (upper right), 2 (lower left), and 4 (lower right).

$$W(t) = L^2(t) \cdot \max(1 - A^2(t), 0)^2 \cdot V(t) \cdot I^2(t) \cdot (1 + 2\tau), \tag{1}$$

which is $(1+2t/T)$ times $W_{f_0}(t)$ from Kochanski *et al.* (2005). In Eq. (1),

- $L(t)$ is the loudness (following Stevens, 1971),
- $A(t)$ is a measure of periodicity (ranging from 0 to approximately 1),
- $V(t)$ is the binary voicing estimate from the pitch detector, and
- $I(t)$ is a manual indicator of irregular voicing (its value is most often 1, but the value occasionally ranges down to zero for regions with no clear pitch), and
- τ is the normalized time, defined below in Sec. II F 2, which ranges from 0 to 1.

The last term in Eq. (1), $(1+2\tau)$, partially compensates for the typical decrease in $L(t)$ as an utterance progresses. Without some compensation, $W(t)$ would generally decline along the sentence. However, in two-accent sentences such as we use, the second (nuclear) accent is believed to be perceptually and semantically at least as important as the first (Cruttenden, 1997). The last term in Eq. (1) was chosen to provide approximate equality in $W(t)$ for the two accents. We have found that our results would not be substantially different if the coefficient in the last term were 1 or 4. Also, within the context of Kochanski *et al.* (2005), changing the weight function by raising it to the power 2 or 0.5 had minor effects. Consequently, we do not expect that the results presented here are critically dependent on the detailed definition of $W(t)$.

2. Normalization

For each f_0 track, we normalized the time axis to range from 0 to 1 by computing $\tau_{i,j}=t_{i,j}/T_i$, where i indicates an utterance, j selects a datum in the utterance, T_i is the duration of the ith utterance, and t and τ refer to real and normalized time, respectively. We normalized f_0 by dividing it by $\bar{f}$, the 10%-trimmed weighted average of $f_0(t)$ over all that subject's sentences from all iterations, and converting the quotient to semitones. This gave a normalized fundamental frequency $\phi(\tau)=12\log_2[f_0(t)/\bar{f}]$.

III. RESULTS AND DISCUSSION

To check that the subjects' productions were representative of English, a phonologist (author BB) applied standard intonation labels (Beckman and Ayers, 1997) to a subset of the data. The labelling and classification of the contours generally proved straightforward. Ninety-six percent of the first imitations resulted in a previously recognized English contour. Furthermore, the contours tended to be stable. A contour received the same annotation in all four iterations 65% of the time; changes between iterations usually yielded another recognized English contour. Overall, only 15% of the contours could not be unambiguously classified.

Figure 3 shows the f_0 contours of the 100 initial stimuli and of iterations 1, 2, and 4 for subject AD. The areas of the dots are proportional to the loudness of speech multiplied by a measure of of periodicity [see Sec. II F 1, Kochanski *et al.* (2005)].

Near the center of the utterances (e.g., $0.2<\tau<0.6$, where τ is the normalized time), the spread is primarily due to the difference between basis contours E and the other two (A and H). The spread at the tail ($\tau>0.9$) contrasts H vs. (A and E). In the very first iteration, the distribution of f_0 contours for AD already develops some structure beyond that for the initial stimuli. Over the succeeding iterations, the broad distribution of f_0 near the centers of the utterances gradually splits into two branches.

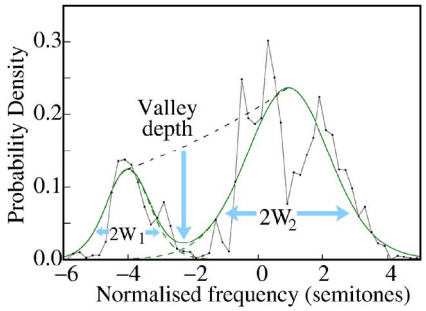

FIG. 4. Fit of a two **t**-density mixture to the central section of iteration 4, subject BP. The abscissa is f_0 in semitones relative to the speaker's average; the ordinate gives probability. The line with dots represents the histogram of the data to which the model was fitted. The smooth black curve is the best-fitting model; the dashed curves show its two components. The vertical arrow indicates where the valley depth is computed, to compare the density minimum to a density interpolated between the peaks of the two components. The width of each component is also indicated.

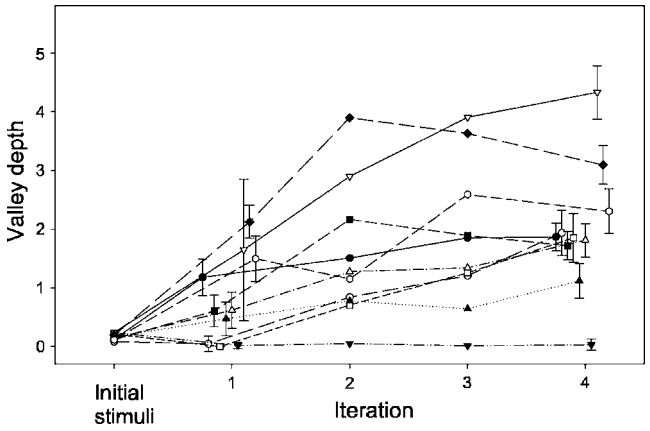

FIG. 5. Measure of bimodality in the distribution of central f_0 for initial stimuli and for mimicry responses as a function of iteration. Each curve gives results for one speaker. Ordinate is valley depth (see Fig. 4). Error bars are 1-σ from Monte Carlo simulations; other error bars (not shown) are of comparable size. (Subjects PC: ▼; EM: ▲: AD: ●; BP: ○; CB: ■; CM: □.)

Most of our subjects manifest clear splitting such as is seen for subject AD. (We show data from subject AD because they are at the median for two measures of this splitting derived below from Fig. 5. AD's final valley depth ranks sixth out of 10, and the increase in valley depth between iterations 1 and 4 ranks sixth out of 10.)

While participants *can* reproduce the intermediate contours [e.g., the "iteration 1" panel in Fig. 3, for $-2<\phi(\tau)<0$, $0.2<\tau<0.6$], those contours are unstable. They move over successive iterations into either the upper or lower stable branch (e.g., "iteration 4" panels of Fig. 3).

A. One branch or two? Modeling the distribution of central f_0

The f_0 tracks in the subjects' productions apparently collapsed into two branches over mimicry sessions. This collapse is consistent with the proposition that a subject's mental structures include certain intonation patterns as attractors. According to the mathematical model, the branches should become arbitrarily narrow, but human variability places a lower limit on the width of the branches. Statistical testing of the splitting of the f_0 tracks in our data therefore becomes necessary. To do this, we take a slice of data for each combination of subject and iteration between $\tau=0.375$ and $\tau=0.5$. Inspection showed that the distribution of such normalized f_0 values has longer tails than would a unimodal model constructed from a single Gaussian density or a bimodal model constructed from two Gaussians. We therefore represent the histogram of $\log(f_0)$ as a mixture of two of Student's **t** probability densities, one for each branch. This allows models that are somewhat heavier-tailed than Gaussian mixtures. Figure 4 shows the histogram of $\log(f_0)$ for iteration 4 from subject BP (lines with dots). It also displays the best-fitting mixture of **t** densities, along with the width of each density. The valley-depth measure is discussed below.

Through a statistical analysis, we compare how well one-and two-component models of the distribution represent such data from each subject and each iteration. The analysis uses a Markov chain Monte Carlo approach based on Bayes' Theorem. It produces samples from the joint distribution of the parameters that define the mixture and calculates the likelihood ratio (relative a posteriori probabilities) of the one- and two-density models for each subject.

1. Models for the probability density of f_0

The one-density model for the probability density of f_0 is

$$H_1(\phi;\theta_1)=\mathbf{t}\left(\frac{\phi-\mu_1}{\sigma_1},\eta_1\right), \tag{2}$$

where θ_1 is shorthand for the three parameters (μ_1,σ_1,η_1), and **t** is Student's **t** function. The two-density model is

$$H_2(\phi;\theta_2)=r\cdot\mathbf{t}\left(\frac{\phi-\mu_{21}}{\sigma_{21}},\eta_2\right)+(1-r)\cdot\mathbf{t}\left(\frac{\phi-\mu_{22}}{\sigma_{22}},\eta_2\right) \tag{3}$$

where θ_2 is shorthand for the five parameters $(\mu_{21},\mu_{22},\sigma_{21},\sigma_{22},\eta_2,r)$. In these equations, μ_k sets the center of a component, σ_k specifies its width, and η_k is the degrees-of-freedom parameter of the **t** density. We are not fitting this density to data. Instead, we are using it as a convenient way to represent a symmetric probability density that is heavy-tailed compared to a Gaussian. Therefore, η_k is a free parameter of the model: it controls how heavy the tails are. For the two-density distribution, r sets the relative probability masses of the two components.

We constrain the parameters so that $\sigma_k>0$, $0<r<1$, and $2<\eta_k<20$. ($\eta\geqslant 20$ makes the **t** density indistinguishable from a Gaussian.) We can safely constrain $\mu_{22}>\mu_{21}$ so that μ_{21} always represents the lower branch and μ_{22} represents the upper branch of f_0, as in the right-hand panels of Fig. 3.

Under the assumption of the one-density model θ_1, the probability of observing the data, $\Phi=\{\phi_{i,j}\}$ for j in the set of utterances and $0.375<\tau_{i,j}<0.5$, is

$$\log[P_1(\Phi|\theta_1)] = \sum_{\Phi} \log[H_1(\phi_{i,j};\theta_1)] \cdot \frac{W_{i,j}}{\bar{W}_j} \cdot \frac{\Delta t}{T_{\text{corr}}}. \quad (4)$$

For the two-density model θ_2, it is

$$\log[P_2(\Phi|\theta_2)] = \sum_{\Phi} \log[H_2(\phi_{i,j};\theta_2)] \cdot \frac{W_{i,j}}{\bar{W}_j} \cdot \cdot \frac{\Delta t}{T_{\text{corr}}}. \quad (5)$$

In these equations, $\bar{W}_j$ is the average weight across an utterance, and $\Delta t/T_{\text{corr}}$ is the ratio of the interval over which f_0 is measured, compared to the assumed correlation time of the f_0 measurements. We conservatively assumed T_{corr} = 100 ms, so $\Delta t/T_{\text{corr}}$=0.1. Effectively, this means that we use only 10% of our data in computing the significance of the two-component model. The actual correlation length of the f_0 measurements obtained from `get_f0` is hard to estimate precisely, due to the complexity of the algorithm. In strongly voiced regions, however, `get_f0` often reflects changes in fundamental frequency that occur over 10–20-ms time scales. A suitable value for T_{corr} therefore may well be substantially smaller than our 100-ms assumption. Reducing T_{corr} would make the two-component model more significant relative to the one-component model.

2. Statistical evaluation of the models

We used Bayes' theorem and a Markov chain Monte Carlo algorithm (Geyer, 1992; Metropolis *et al.*, 1953) to generate samples from the distributions of $P_1(\theta_1|\Phi)$ and $P_2(\theta_2|\Phi)$. These are the posterior joint probability densities of the parameters, given the observed data. We assumed flat priors over the parameter ranges specified above.

To implement statistical testing, we computed the likelihood ratio of two classes of hypotheses, where each class consists of a set of related models:

(1) the class of single-component distributions that are plausible fits to the data [specifically Eq. (2) where θ_1 selects the model within the class] and
(2) the class of two-component distributions that are plausible fits to the observed data [specifically Eq. (3) where θ_2 selects the model].

This is a straightforward application of Bayesian model averaging (Hoeting *et al.*, 1999). Assuming flat prior probabilities and no bias toward either model, the likelihood ratio is

$$R = \frac{\langle P_1(\theta_1|\Phi)\rangle}{\langle P_2(\theta_2|\Phi)\rangle} = \frac{\langle P_1(\Phi|\theta_1)\rangle}{\langle P_2(\Phi|\theta_2)\rangle}, \quad (6)$$

where the angle brackets, $\langle\ \rangle$, denote an expectation over the corresponding probability density, e.g.,

$$\langle P_1(\Phi|\theta_1)\rangle = \int P_1(\Phi|\theta_1) P_1(\theta_1|\Phi) d\theta_1. \quad (7)$$

The expectation value can therefore be implemented as an average over samples generated by the Markov chain Monte Carlo process, because the process picks samples from each region of volume $d\theta_1$ with probability $P_1(\theta_1|\Phi)\cdot d\theta_1$. Therefore,

$$\langle P_1(\Phi|\theta_1)\rangle = \sum\nolimits_{\theta_1} P_1(\Phi|\theta_1) \Big/ \sum\nolimits_{\theta_1} 1. \quad (8)$$

The expectation value for $\langle P_2(\Phi|\theta_1)\rangle$ is computed similarly.

If the likelihood ratio $R<1$, then the two-component model is a better representation of the data. By the rules of hypothesis testing, however, we do not reject the single-component model unless R falls below a confidence limit of 0.001. Our tests show that the two-density model for iterations 2–4 is statistically significant (with R actually less than 10^{-4}) for all but subject PC. On iteration 4 for PC, R <0.01 and the means of the components differ by 2.2 semitones. The components, however, have widths of 1.7 and 0.9 semitones and therefore overlap seriously. Although PC's distribution of f_0 values may require two components to represent it well, the components do not separate into clearly distinct branches.

In summary, in almost all cases, two **t** densities give a significantly better fit to the observed distributions of f_0 than does one. The best-fit **t** densities typically have noticeably but not dramatically longer tails than would Gaussians. We find η=13 with an intersubject standard deviation of 5 in Eqs. (2) and (3).

B. Valley depth

Having two components is a necessary but not sufficient condition for establishing bimodality, as the **t** densities of a two-component fit could overlap strongly. Their means could even virtually coincide. To provide further evidence for bimodality, we compute a valley depth measure (Fig. 4) from a fit of two mixed **t** densities to a histogram of central f_0. The computation finds the minimum of the curve of $H_2(\phi;\theta_2)$ between μ_{21} and μ_{22} and compares it to the value interpolated between the peaks. The minimum is obtained by simple iterative searching; the interpolation is linear on $\log(H_2)$ between $H_2(\mu_{21};\theta_2)$ and $H_2(\mu_{22};\theta_2)$. The valley depth is the negative log of the density at the minimum, divided by the interpolated density at the same point. Valley depth is zero or negative for a unimodal density. Values greater than one typically imply two well-separated components and hence two clusters.

The valley depth measure was obtained for each Monte Carlo sample of θ_2 generated from $P_2(\theta_2|\Phi)$. Figure 5 shows the set of valley depths for each speaker plotted against iteration number. The lines show the mean; standard deviations are shown for the first and last iterations. The error bars are conservative. As stated previously, our computations assumed that f_0 measurements are correlated over a large 100-ms stretch because of the dynamic programming algorithm in `get_f0`. The correlation length in such strongly voiced regions, however, is typically no larger than 20 ms. The error bars in the figure would shrink in proportion to the square-root of the correlation length. They are perhaps about half as large as plotted. The figure also shows fits to the f_0 contours for the initial stimuli.

The mean valley depth starts at 0.16 for the stimuli. Averaged over all subjects, it increases from 0.84 for itera-

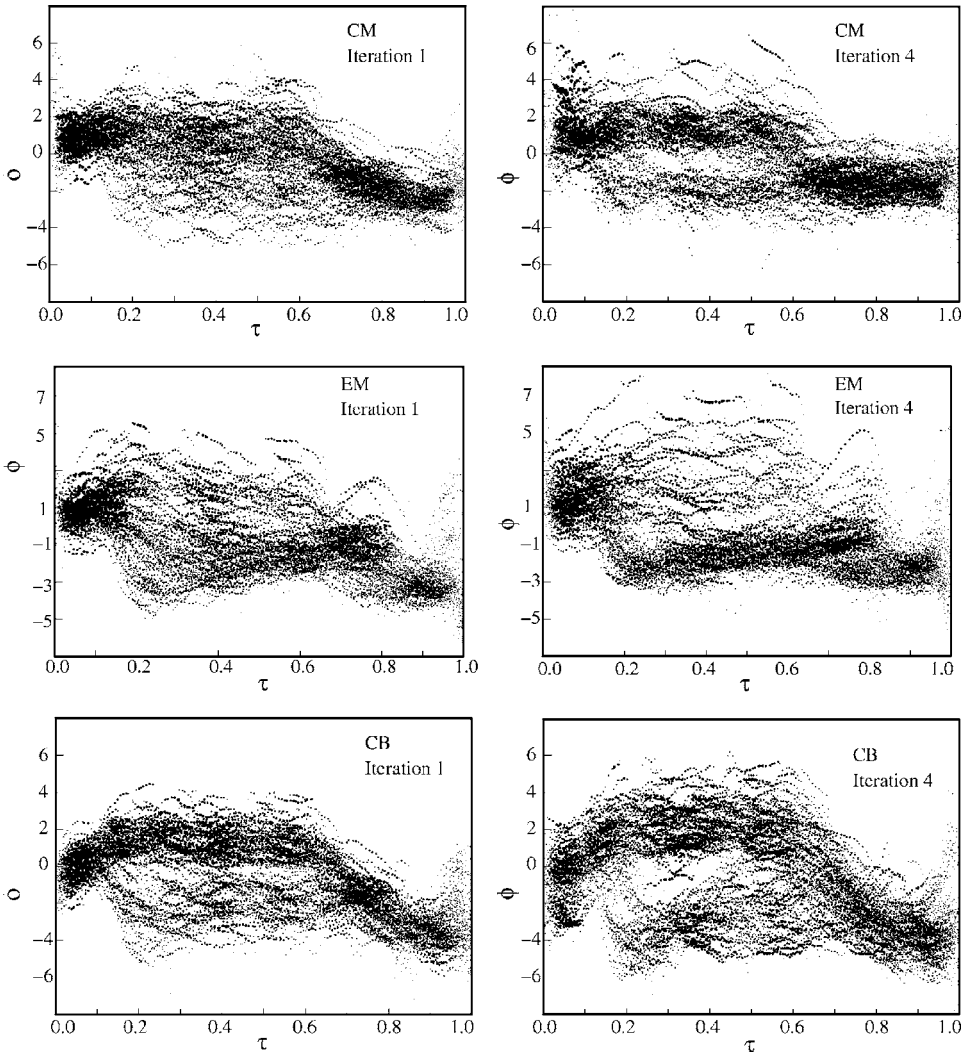

FIG. 6. Variation between subjects. This figure shows normalized f_0 contours, $\phi(\tau)$, for subjects CM, EM, and CB (top to bottom), plotted as per Fig. 3. Data from iterations 1 (left) and 4 (right); the initial stimuli are identical, as displayed in the upper-left panel of Fig. 3.

tion 1 to 1.96 for iteration 4. This is a highly significant increase ($z=8.2$, $P<10^{-6}$). Individually, nine subjects show some increase, and seven of them show a significant increase ($P<0.05$).

The exceptional subject PC (point-down filled triangles) gives a valley depth close to zero in most of the Monte Carlo samples. There is no valley between the two components fitted to PC's data, in line with the relatively high values of R found above for that subject.

Most subjects take more than one iteration to reach a stable state. Their f_0 contours did not collapse immediately into two branches. Therefore, between perception and production on a given trial, most subjects remember some details of the stimulus beyond the shape of a particular "stable attractor" contour. Intonational phonology with its choices among a small set of discrete contours cannot be a complete description of the subject's mental representations of the contours they hear.

C. Variation in and around the attractors

Subjects do not finally settle exactly on the attractor contours. The standard deviations of the two branches shrink (see Figs. 3 and 4) to about 0.7 semitones for the lower branch and 1.1 semitones for the upper. Frequency shifts smaller than this may often not be linguistically useful because they cannot be reliably perceived, remembered, or reproduced. These standard deviations are similar to rms differences between repetitions of the same utterance (Holm and Bailly, 2002, Sec. III A).

Figure 6 shows the variation in attractors across subjects. Generally, the upper branch either rises early or starts high, stays high, and then drops. The lower branch either starts high or has an early peak near $\tau=0.1$ and typically has a smaller peak around $\tau=0.8$. Final rises are common but often they sound flat or even falling, perhaps due to the final decline in loudness starting near $\tau=0.92$.

Subject CM (top of Fig. 6) has no peaks in the lower branch (unlike AD in Fig. 3). No bimodality is visible in iteration 1, but branches develop in iteration 2 or 3 (not shown) and are clear by iteration 4. Subject EM (middle) has a poorly defined upper branch, with a valley depth that is next-to-lowest in Fig. 5. Even for this subject, however, a lower branch forms by iteration 1 and becomes narrower between iterations 1 and 4. Subject CB (bottom panel) shows two branches even in iteration 1 which spread apart over the iterations. CB has a well-defined peak early in the bottom branch, near $\tau=0.15$, and a broad, ill-defined later peak.

The initial stimuli include both rising-tail and falling-tail

f_0 tracks, derived from the three basis contours. Nonetheless, we do not clearly observe independent traces of the third basis contour in the productions, nor do we see strong distinctions in the tails of the productions. Only one subject (AD, shown in Fig. 3) has three clear groups of contours. While several subjects show late bimodality around $\tau=0.95$, correlations between the tails and the centers of utterances complicate any interpretation.

In summary, the f_0 contours are typically bimodal. They become more strongly bimodal with every successive iteration. We see attractors for two of the three basis contours that we expected; overall our data do not show an association of contour H (rising tail) with an attractor.

D. Qualitative analysis

In the phonologist's (author BB) labeling three contours occurred particularly often: two peaks (35%), a hat pattern (40%), and one with a final rise (17%). Recall that 96% of the labels applied to the first imitations resulted in a previously recognized English contour and that the labels showed stability over iterations. The labeling therefore indicates a collapse in the very first iteration towards some stable attractor, whereas the acoustic analysis points to a gradual movement towards it. Linguistic categories could be assigned to most of the contours, even if—acoustically—a contour is not yet close to a stable attractor but is within its basin of attraction. This could be interpreted as a "magnet effect" for English intonation, analogous to Kuhl (1991) and Guenther and Gjaja (1996).

One interpretation of this result is that the attraction happens at the perceptual level rather than at the level of motor control. If so, then the phonologist perceived the speech by way of a mapping that is presumably similar to the subject's. The phonologist's perception of a contour from iteration 1 would then have undergone two mappings: one imposed by the subject on an initial stimulus and one imposed by the phonologist on the subject's response. The phonologist's perception would then effectively be one iteration further along than the acoustic analysis. Her perceived contours therefore would be closer to the (discrete) attractors.

Our findings are compatible with the results of Ladd and Morton (1997). They reported that subjects discriminate between numerous contours that (for instance) signal presence of emphasis but that subjects also readily assign contours to either an "emphasis" or "no emphasis" category. The corresponding feature of our results is the preservation of some of the difference information, as evidenced by the gradual collapse to the attractor.

One speculative way to explain the ease of conscious classification, the slow collapse of the acoustic properties toward attractors, and the discrimination of contours within categories is to assume that a heard f_0 contour is internally represented as an attractor plus a partially remembered difference between that attractor and the presentation. The phonologist may then suppress the difference information when classifying contours, retaining only the category information. We recognize, however, that there are other possible explanations of the three effects enumerated above.

IV. CONCLUSIONS

Although subjects can hear and imitate randomly generated contours that are not normally found in English, over several iterations their productions converge onto a limited set of distinct contours. These function mathematically as attractors and correspond to some (but not necessarily all) common English intonation contours. The results of iterative mimicry provide objective support for the existence of basic intonation patterns that act as attractors. Plausibly, the attractors are either a description of clusters of episodic traces/exemplars (Goldinger, 1998), or are a soft targets along the lines of (e.g., Kochanski *et al.*, 2003) that represent the position of the cluster center but allow some variability around it. Our results are not consistent with the hypothesis that subjects have available only a discrete phonological description of the intonation. Subjects actually perceive, remember, and use acoustic detail above and beyond what international phonology normally represents.

Our attractor contours have a parallel in the production and perception of vowels. In speech production, vowels are highly variable. Nevertheless, infants extract the vowels of their native language, and the adult vowel space contains regions of substantial size where a given vowel is heard reliably. The Perceptual Magnet Theory of Phonetic Learning (Guenther and Gjaja, 1996; Kuhl, 1991) shows how these local vowel regions emerge: a set of magnets acts as a prototype and warps the continuously variable vowel space around each of them.

While there are similarities with vowel magnets, intonation contours are dramatically extended in time. Vowels can be identified from brief bursts of sound (20–60 ms), while the attractors seen here are global properties of an entire utterance (about 1 s). Consequently, the mechanism that recognizes and processes these extended attractors may be different from the mechanism that handles vowels.

Stable attractors in the mapping between produced intonation and heard intonation have implications for language learning and development. To the extent that mimicry is involved in the normal use of language, the distribution of intonation contours that people hear should have a high density of examples near each attractor. Notably, children should be exposed to this distribution of contours. If the attractor contours in the children's mappings are based upon especially commonly heard examples, this could provide the mechanism for the stable transmission of the international aspects of language from one generation to the next in analogy to Maye *et al.* (2001) and Saffran *et al.* (1996).

ACKNOWLEDGMENTS

This work has been supported by the Oxford University Research Development Fund. We thank John Coleman for assistance with the stimulus generation and for tolerance and advice, and we thank Cindy Pribble Kochanski for helping to test the experimental software and design.

APPENDIX: SYNTHESIS PROCEDURES

A LPC analysis was made of each of the utterances from the original male and female speakers, using the Entropics

program `refcof`). Each analysis yielded 17 reflection coefficients; three voice source parameters were also obtained from `get_f0`.

For each of the candidate utterances, the parameters for the versions with falling and rising intonation were matched by dynamic time warping. This algorithm finds a monotonic mapping between the time axis of one utterance and the time axis of another that minimizes the mean magnitude of the vector difference between two sets of parameters. This yielded a time series of the 20 speech parameters that was half-way between that of the two starting utterances, both in terms of segment durations and of segmental properties like formant frequencies. Each of the candidate sentences was then synthesized from the intermediate parameter.

Bartlett, F. C. (**1932**). *Remembering* (Cambridge U. P., Cambridge).

Beckman, M., and Ayers, G. (1997). "Guidelines for ToBI labelling," Tech. Report, Linguistics Department, Ohio State University, http://ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf.

Brazil, D. (**1985**). *The Communicative Value of Intonation in English*, Discourse Analysis Monograph, No. 2 (Bleak House Books and English Language Research, Birmingham, UK).

Cruttenden, A. (**1997**). *Intonation*, 2nd ed. (Cambridge U. P., Cambridge).

Doll, J. J., and Thomas, D. R. (**1967**). "Effects of discrimination training on stimulus generalization for human subjects," J. Exp. Psychol. **75**, 508–512.

Eisner, F., and McQueen, J. M. (**2005**). "The specificity of perceptual learning in speech processing," Percept. Psychophys. **67**(2), 224–238.

Francis, A. L., Ciocca, V., and Ng, B. K. C. (**2003**). "On the (non)categorical perception of lexical tones," Percept. Psychophys. **65**, 1029–1044.

Geyer, C. J. (**1992**). "Practical Markov Chain Monte Carlo," Stat. Sci., 473–483.

Goldinger, S. D. (**1998**). "Echoes of echoes? An episodic theory of lexical access," Psychol. Rev. **105**, 251–279.

Grabe, E., Rosner, B. S., García-Albea, J. E., and Zhou, X. (**2003**). "Perception of English intonation by English, Spanish, and Chinese listeners," Lang Speech **46**(4), 375–401.

Guenther, F. H., and Gjaja, M. N. (**1996**). "The perceptual magnet effect as an emergent property in neural map formation," J. Acoust. Soc. Am. **100**, 1111–1121.

Gussenhoven, C., and Rietveld, A. M. C. (**1997**). "Empirical evidence for the contrast between H^* and L^* in Dutch rising intonation contours," in *Intonation: Theory, Models and Applications, Proceedings of an ESCA Workshop*, edited by A. Botinis, G. Kouroupetroglou, and G. Carayiannis.

Hirst, D., and Cristo, A. D., (Eds.) (**1998**). *Intonation Systems: A Survey of Twenty Languages* (Cambridge U. P., Cambridge, UK).

Hoeting, J. A., Madigan, D., Raferty, A., and Volinsky, C. T. (**1999**). "Bayesian Model averaging: A tutorial," Stat. Sci. **14**(4), 382–417; http://www.stat.colostate.edu/~jah/papers/statsci.pdf.

Holm, B., and Bailly, G. (**2002**). "Learning the hidden structure of intonation: Implementing various functions of prosody," in *Proceedings of Speech Prosody 2002*, http://www.isca-speech.org/archive/sp2002/sp02_399.pdf.

Jun, S.-A., (Ed.) (**2005**). *Prosodic Typology. The Phonology of Intonation and Phrasing* (Oxford U. P., Oxford, UK).

Kingdon, R. (**1958**). *The Groundwork of English Intonation* (Longman, London).

Kochanski, G., Shih, C., and Jing, H. (**2003**). "Quantitative measurement of prosodic strength in Mandarin," Speech Commun. **41**(4), 625–645; http://dx.doi.org/10.1016/S0167-6393(03)00100-6.

Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (**2005**). "Loudness predicts prominence: Fundamental frequency lends little," J. Acoust. Soc. Am. **118**(2), 1038–1054; URL http://dx.doi.org/10.1121/1.1923349.

Kohler, K. J. (**1991**). "Terminal intonation patterns in single-accent utterances of German: Phonetics, phonology and semantics," Arb. Inst. Phon. Univ. Kiel (AIPUK) **25**, 115–185.

Kuhl, P. (**1991**). "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories," Percept. Psychophys. **50**, 93–107.

Ladd, D. R., and Morton, R. (**1997**). "The perception of intonational emphasis: continuous or categorical?," J. Phonetics **25**, 313–342.

Ladefoged, P. (**1989**). "A note on 'Information conveyed by vowels'," J. Acoust. Soc. Am. **85**, 2223–2224.

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (**1957**). "The discrimination of speech sounds within and across phoneme boundaries," J. Exp. Psychol. **61**, 379–388.

Loeb, D. F., and Allen, G. D. (**1993**). "Preschoolers' imitation of intonation countours," J. Speech Hear. Res. **36**(1), 4–13.

Maye, J., Werker, J. F., and Gerken, L. (**2001**). "Infant sensitivity to distributional information can affect phonetic discrimination," Cognition **82**, B101–B111.

Meltzoff, A. N., and Decety, J. (**2003**). "What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience," Philos. Trans. R. Soc. London **248**, 491–500.

Metropolis, N., Rosenbluth, A. E., Rosenbluth, M. N., Teller, A. H., and Teller, E. (**1953**). "Equations of state calculations by fast computing machines," J. Chem. Phys. **21**, 1087–1092.

O'Connor, J. D., and Arnold, G. F. (**1961**). *Intonation of Colloquial English* (Longman, London).

Pierrehumbert, J. B., and Hirschberg, J. (**1990**). "The meaning of intonation in the interpretation of discourse," in *Intentions in Communication*, edited by P. Cohen, J. Morgan, and M. Pollack (MIT, Cambridge, MA), Chap. 14, pp. 271–311.

Pierrehumbert, J. B., and Steele, S. A. (**1989**). "Categories of tonal alignment in English," Phonetica **46**, 181–196.

Plomp, R. (**2002**). *The Intelligent Ear: On the Nature of Sound Perception* (Erlbaum, Mahwah, NJ).

Remijsen, B., and van Heuven, V. J. (**1999**). "Gradient and categorical pitch dimensions in Dutch: Diagnostic test," in *Proc. of the 14th International Congress of the Phonetic Sciences (ICPhS)*, San Francisco, CA, pp. 1865–1868.

Repp, B. H., and Williams, D. R. (**1987**). "Categorical tendencies in self-imitating self-produced vowels," Speech Commun. **6**, 1–14.

Ryalls, J., Dorze, C. L., Lever, N., Quellet, L., and Larfeuil, C. (**1994**). "The effects of age and sex on speech intonation and duration for matched statements and questions in French," J. Acoust. Soc. Am. **95**(4), 2274–2276.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (**1996**). "Statistical learning by 8-month-old infants," Science **274**, 1926–1928.

Scherer, K. R. (**1985**). "Vocal affect signalling: A comparative approach," in *Advances in the Study of Behavior*, edited by J. Rosenblatt, C. Beer, M.-C. Busnel, and P. Slater (Academic, New York), Vol. **15**, pp. 189–244.

Schouten, B., Gerrits, E., and von Hessen, A. (**2003**). "The end of categorical perception as we know it," Speech Commun. **41**, 71–80; http://dx.doi.org/10.1016/S0167-6393(02)00094-8.

Snow, D. (**1998**). "Children's imitation of intonation contours: Are rising tones more difficult than falling tones?," J. Speech Lang. Hear. Res. **41**(3), 576–587.

Steedman, M. (**2000**). "Information structure and the syntax-phonology interface," Linguist. Inq. **31**(4), 649–689.

Stevens, S. S. (**1971**). "Perceived level of noise by mark VII and decibels," J. Acoust. Soc. Am. **51**(2), (part 2) 575–602.

Talkin, D. (**1995**). "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, edited by W. B. Klein and K. K. Palival (Elsevier, Amsterdam).

Vallduvi, E., and Engdahl, E. (**1996**). "The linguistic realisation of information packaging," Linguistics **34**, 459–519.