INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Perception of Pitch Contours in Speech and Nonspeech

Daniel R. Turner[1], *Ann R. Bradlow*[1], *Jennifer S. Cole*[1]

[1]Northwestern University
dturner@u.northwestern.edu

Abstract

The pitch perception literature has been largely built on experimental data collected using nonspeech stimuli, which has then been generalized to speech. In the present study, we compare the perceptibility of identical pitch movements in speech and nonspeech that vary in duration and in pitch range. Our nonspeech results closely replicate earlier findings and we show that speech is a significantly more difficult medium for pitch discrimination. Pitch movements in speech have to be larger and longer to achieve the salience of the most common speech analog, pulse trains. The direction of pitch movement also affects one's ability to discern pitch; in particular falling excursions are the most difficult. We found that the perceptual threshold for falling pitch in speech was more than 100 times that of previous estimates with nonspeech stimuli. Our findings show that the perceptual response to nonspeech does not adequately map onto speech, and future work in speech research and its applications should use speech-like stimuli, rather than convenient substitutes like pulse trains, pure tones, or isolated vowels.

Index Terms: pitch perception, speech perception, speech synthesis, speech resynthesis, just noticeable differences

1. Introduction

The human ability to perceive pitch patterns in speech has, with few exceptions, been quantified using stimuli that do not resemble human speech. Most research in this area has tested listener perception of pitch using nonspeech carrier signals like pure tones or pulse trains, with the goal of comparing responses between populations, such as native speakers of tonal languages or professional musicians versus speakers of non-tonal language and non-musicians. The present study seeks to validate—or update—the conventional wisdom on the basic facts of pitch perception in speech, and to shed light on what properties of speech best predict perceptibility of the direction of pitch change. Our findings are particularly informative for speech researchers and engineers whose work entails that listeners comprehend pitch patterns. They are also relevant for basic understanding of auditory psychoacoustics and the neural encoding of pitch.

2. Background

Early advances in understanding pitch perception were driven by clinically-oriented audiometry and engineering challenges from the telephone industry in the 1950s. Using an analog formant synthesizer and pulse train generator, [4] estimated the just-noticeable difference (JND) between two nondynamic stimuli to be about 0.3Hz using a simple pairwise discrimination task. Pulse trains roughly resemble the human glottal waveform and are considered to be more speech-like than pure tones, another common spoken pitch substitute. By convention, JND is defined as the point where listener accuracy exceeds the arbitrary threshold of 75%. [6] replicated the lower bound found by [4], but only for a sustained /ɛ/ vowel administered at 120Hz; for an /ɛ/ vowel on a steeply falling linear ramp the reported threshold was 4Hz, or about 13 times greater. A more modern treatment of the problem came from [2] which compared native Mandarin to native English speakers in their perception of pulse trains. If we apply the conventional JND threshold to findings in [2], it suggests a JND of 25Hz, about 6 times that of [6]. These results are difficult to reconcile and many factors could have contributed to the differences between the findings of [2] and earlier work: stimuli in [2] are almost twice as long (400ms versus 250ms), they explore much higher frequencies (250Hz or 300Hz versus 120Hz), and the participants had three response options rather than two. Since the JND threshold is fixed at 75%, the number of response options drastically changes the difference between chance accuracy and the JND threshold. Recently, [5] reported a study that used resynthesized human-produced /a/ vowels to compare discrimination of pitch direction, height, and slope in native Mandarin and native English listeners. They found a JND of about 7Hz for both language groups, but significantly different response patterns to pitch slope and pitch height. Native Mandarin speakers were more sensitive to pitch slope and native English speakers were more sensitive to pitch height. While the original vowels used to create stimuli in [5] were produced by a human, natural speech is much more complex than isolated vowels and this may impact the ability to discriminate pitch information, such as its direction.

In sum, while previous studies that used nonspeech or isolated resynthesized vowels have been crucial for understanding the absolute envelope of perception in idealized signals, it is unclear how their findings generalize to the natural task and signal of comprehending pitch information in human speech. The present work frames these concerns with the hypothesis that the complexity of speech makes discrimination of embedded pitch information more difficult than nonspeech. Natural human speech is produced with concurrent segmental and suprasegmental content including rich spectral information, and the listener must parse multiple parts of the acoustic stream to understand its meaning. Unlike the task of discriminating which of two pulse trains has the higher pitch, interpreting natural language often involves making more fine-grained judgements about details of the pitch contour, such as its direction. Pitch direction can signal many differences in meaning, from sentence type to speaker affect. Therefore, not only is speech a more complex signal for listeners, but the task of judging the direction of pitch in speech is more complex compared to typical tasks from the JND literature. The goals of the present research are to compare the perception of nonspeech to speech and discover what properties of pitch movements (size, duration, etc.) best predict accurate perception of pitch

http://dx.doi.org/10.21437/Interspeech.2019-2619

content. To achieve this, we present the following experiment wherein participants categorize a large set of pitch contours in speech and nonspeech as rising, falling, or flat.

3. Methods

3.1. Task

Participants classified the pitch in each stimulus as rising, falling, or flat as quickly as possible.

3.2. Stimuli

To test the hypothesis that pitch perception is more difficult in complex, speech-like signals, this study compares pitch identification in speech and nonspeech. Identical to [2], our nonspeech condition consists of computer-generated pulse trains created in Praat using the "Create sound from tone complex" function [3]. Our speech condition consists of recordings from a trained speaker producing English-like nonce words. These were designed to optimize pitch transmission, counterbalance for implicit pitch of vowels, and maintain word medial stress. Recordings were resynthesized using the Time-Domain Pitch-Synchronous Overlap-and-Add method (PSOLA) method, identical to [5]. This method effectively interpolates a pulse train with the natural recordings, so the pitch contour of the speech and nonspeech conditions are identical. Pitch excursions followed the same inventory employed by [2], shown in Table 1:

Table 1: *Pitch excursion inventory*

Direction	Onset	Change
Rising	250Hz or 300Hz	5, 10, 15, 20, 30, 40, or 50Hz
Falling	250Hz or 300Hz	5, 10, 15, 20, 30, 40, or 50Hz
Flat	200, 210, 220, 230, 240, 250, 260, 270, 280, 300, 320, 330, 340, or 350Hz	0Hz

Stimuli also varied in duration (400ms or 1100ms) primarily to explore the role of pitch slope in the categorization of stimuli for the direction of pitch movement in the two stimulus types. In the speech condition, the two durations corresponded to 1- and 3-syllable nonce words that conform to English phonotactics and stress patterns, such as *ba* in *bazagi*. The first syllable in the trisyllabic nonce words began with /b/, /d/, or /g/ followed by /ə/; the second syllable began with /m/, /n/, /v/, /z/, /l/, or /w/ followed by /i/ or /a/; the third syllable began like the first (/b/, /d/, /g/) and ended with /i/ or /a/, alternating with the vowel of the second syllable. From this formula, two sets of 18 trisyllabic nonce words were randomly selected to serve as the test and practice sets. Half of the words had /a/ in the second syllable and /i/ in the third syllable, and half had the reverse order. 18 monosyllable nonce words were similarly constructed from the same consonant set, with vowel /i/ or /a/. In the written instructions for the experiment, words were presented orthographically with "ee" representing /i/ and "a" for /a/.

3.3. Participants

32 right-handed native English speakers were recruited from undergraduate linguistics courses at Northwestern University and were given course credit for their participation. Left-handed participants were excluded because we wanted to analyze response time between participants for keystroke responses, and handedness is known to influence response latency. Participants who reported speech, hearing, or reading problems (n=2), or who fell below a threshold of 50% accuracy on comprehension questions (n=1) were excluded and replaced. We did not analyze participant survey responses except to determine eligibility.

3.4. Procedure

After reviewing and signing a consent form, participants completed a brief background survey before the experiment began. The first block of the experiment was an interactive tutorial which oriented participants to the task, controls, and types of stimuli. Participants were assigned to one of four counterbalanced groups which determined the order of experimental blocks. Stimuli were presented in four blocks, alternating in the type of stimuli (speech and nonspeech) and duration (short and long). Trials within each block were fully randomized. Every block began with a practice set of 18 trials, with accuracy feedback following each trial. After the 18 practice trials, a test set consisting of 3 repetitions of 42 pitch excursions began, yielding a total of 168 critical trials per block, for 672 critical trials per participant. 10% of the trials, selected at random, were immediately followed by a comprehension question (for speech) or an instruction to press a specified response key (for nonspeech). Comprehension questions asked the participant if the onscreen word matched the preceding trial's stimulus. Half of the comprehension questions matched the prior trial, while the other half showed a random competitor from another trial. For nonspeech, the directed key presses helped us to screen for potentially inattentive participants. An inter-stimulus interval of 2100ms was chosen for all trials based on feedback from pilot participants. The experiment was administered using PsychoPy3 running on an Apple Mac Mini in a sound-attenuated booth in the Northwestern Phonetics Laboratory. Participants used a standard keyboard to indicate the perceived direction of pitch: right shift for rising, left shift for falling, and space for flat. The experiment took 50 minutes to complete on average.

3.5. Statistical Model

While our study design partly replicates [2], we deviate from all previously mentioned work in our use of a generalized linear mixed effects regression model to fit our data. Every critical trial was coded for reaction time and accuracy but, because we had no predictions regarding latency, we chose to create a comprehensive accuracy model: Accuracy ~ (Type + Depart + Duration + Change)^4 + (Type + Depart + Duration + Direction)^4 + (1 | participant). Our dependent variable was the binomial trial accuracy, 1 for correct or 0 for incorrect, which was interpreted with a logit link. Fixed effects included type (speech/nonspeech), direction (rise/fall/flat), change (in scaled semitones), duration (in scaled milliseconds), and the starting point of the pitch contour ('depart', in scaled Hertz) and the model included random effects by participant. The model included all 2-, 3-, and 4-way interactions, with the exception of interactions that intersected direction and change, as flat contours do not change in pitch. Modeling was done using the lme4 package in the statistical software R [1, 8].

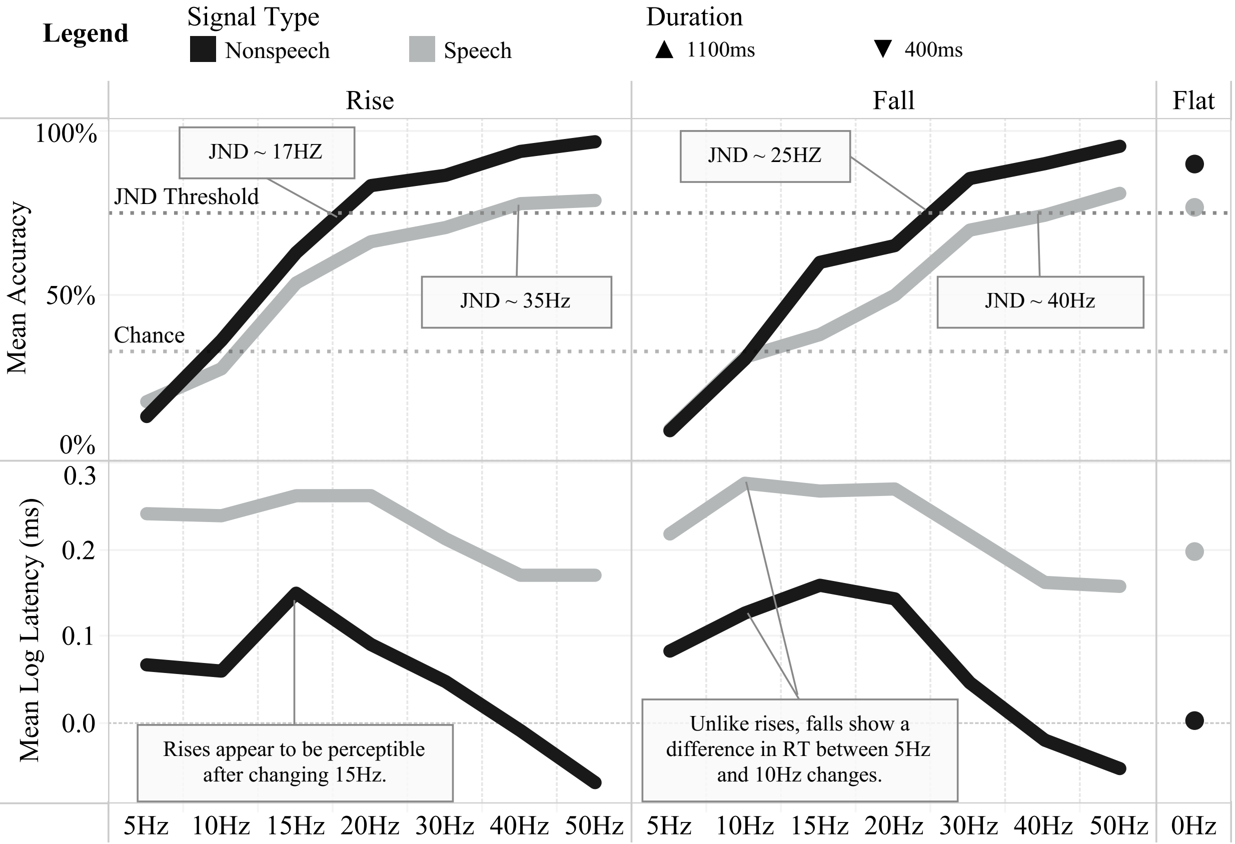

Figure 1: *Average trial accuracy and latency by amount of change, sorted by type and direction*

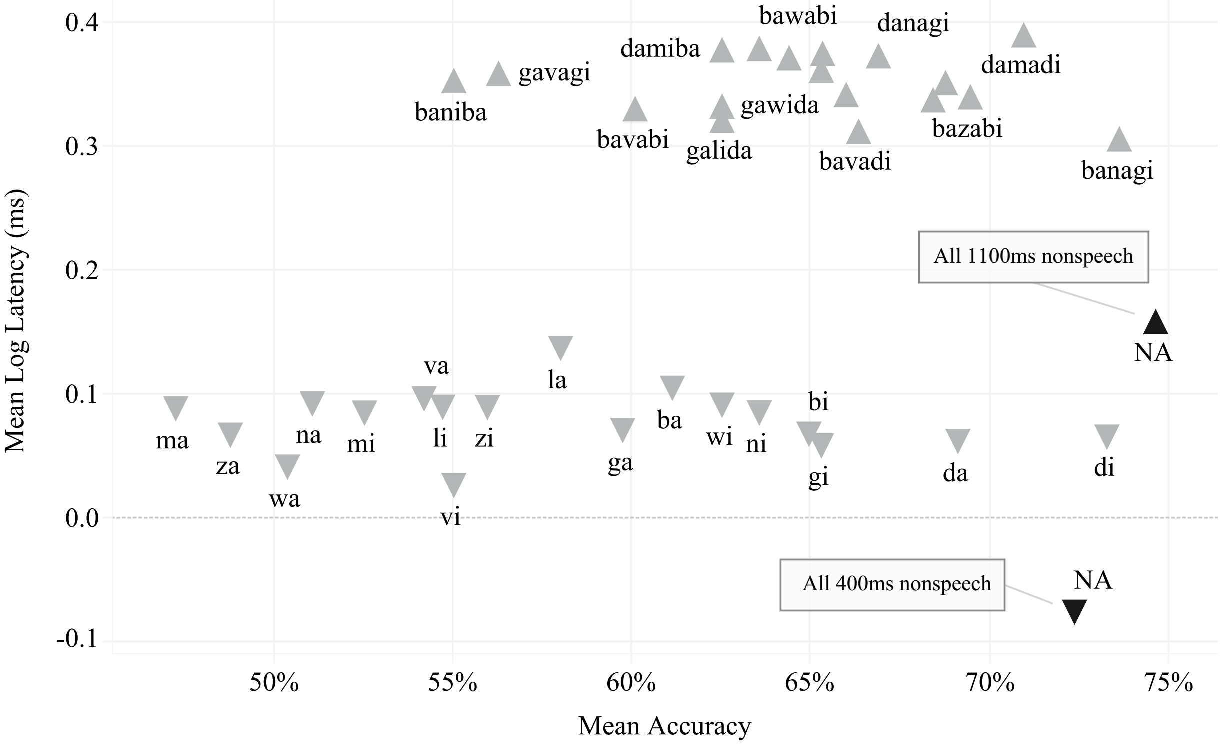

Figure 2: *Average trial accuracy and latency by nonce word, grouped by duration.*

3.6. Predictions

We predicted perceptual facilitation in higher accuracy and/or faster response times for (1) larger pitch excursions compared to smaller excursions; (2) pitch movements in nonspeech compared to speech; (3) pitch movements over longer durations compared to shorter durations.

3.7. Reproducibility

All stimuli, generation and resynthesis code, experimental files, raw data, and our statistical model are available online free of charge through our preregistration page at: https://osf.io/umq9j/

4. Results

Plots of our raw accuracy data show a higher JND for pitch movements in speech than in nonspeech and a higher JND for falling pitch than rises (Figure 1, top panel). Plots of response latency show responses to speech took significantly longer than nonspeech (Figure 1, bottom panel). JNDs were evaluated using the psychophysics convention of fitting response curves to a 75% accuracy cutoff. This yielded a JND of 17Hz for nonspeech rises and 25Hz for nonspeech falls, which contrasts with the JND of 35Hz for speech rises and 40Hz for speech falls. Overall, identification of flat pitch in speech was about 75% accurate while flat pitch in nonspeech was closer to 95% accurate, which we consider ceiling performance. Plots of the reaction time data show that speech takes about 200ms longer to respond to than nonspeech and they suggest that the perceptual threshold for identification of falling contours is lower than for rising, (See Figure 1, bottom pane). While we closely controlled the segmental content of the nonce word stimuli, we did not predict there to be an effect of segmental content on accuracy. We were therefore surprised to find that stop-initial monosyllabic words, like /di/ and /da/, were classified more accurately than continuant-initial ones, like /ma/ and /za/. We did not find a general pattern in response time to individual nonce words, but there are notable differences in accuracy (See Figure 2). Because we did not design nonce words to explore the role of segmental phonetics, we leave the explanation of these and other segmental effects for future research.

Our regression model found all main effects except duration to be highly significant ($p < .001$). Estimated p-values were obtained from asymptotic Wald tests and associated z-values confirm these strong trends. A main effect of type showed pitch identification in speech to be less accurate than nonspeech, confirming patterns we observed in the empirical data ($\beta = -0.733$, se = 0.90, $z = -8.12$, $p < .001$) and a main effect of direction showed that flat contours were more accurately classified than falling ($\beta = 4.57$, se = 0.13, $z = 34.4$, $p < 0.001$) and rising contours were slightly more accurate than falling ($\beta = 0.29$, se = 0.92, $z = 3.18$, $p < 0.002$). We also found a main effect for the start point of pitch contours, ‘depart’, which showed that stimuli with lower starting points were more accurately identified than higher ones ($\beta = -0.46$, se = 0.09, $z = -5.00$, $p < 0.001$). While duration was not found to be significant on its own, it was found to interact significantly with stimulus type—longer speech stimuli were more accurately identified than shorter nonspeech stimuli ($\beta = 0.33$, se = 0.13, $z = 2.57$, $p < 0.02$)—and longer rises were more accurate overall ($\beta = 0.41$, se = 0.13, $z = 3.14$, $p < 0.002$). Because duration also modulates pitch slope in our stimuli, we take from this that our participants attended less to pitch slope than other factors, like pitch height, supporting the conclusions of [5]. Interestingly, speech benefited less from larger pitch movements versus nonspeech ($\beta = -1.0$, se = 0.1, $z = -10.5$, $p < 0.001$) and flat speech stimuli were less accurately identified overall ($\beta = -1.3$, se = 0.17, $z = -7.91$, $p < 0.001$). Table 2 reports all main effects and all significant interactions from the model.

Table 2: *Regression model results*

Predictor	Estimate	se	z	p
Intercept	-0.013	0.142	-0.094	.920
Speech vs. Nonspeech	-0.733	0.090	-8.116	< .001
Change	2.382	0.084	28.31	< .001
Depart	-0.462	0.092	-5.003	< .001
Long vs. Short	n.s.	n.s.	n.s.	n.s.
Flat vs. Fall	4.566	0.133	34.42	< .001
Rise vs. Fall	0.293	0.092	3.183	< .002
Speech:Long	0.327	0.127	2.571	< .020
Speech:Change	-1.029	0.098	-10.49	< .001
Speech:Flat	-1.344	0.170	-7.908	< .001
Depart:Rise	1.071	0.129	8.267	< .001
Long:Rise	0.414	0.131	3.145	< .002

5. Discussion

The present study informs the pitch perception literature in three important ways. First, it establishes that pitch direction in speech is significantly more difficult to classify than in its most common research analogs, pulse trains and pure tones. In fact, pitch contour categorization accuracy for speech benefited less from larger pitch excursions than nonspeech, all else being equal. Second, data from our nonspeech condition almost exactly replicate [2] and our statistical model replicates the conclusion of [5] that native English listeners may attend more to pitch height versus pitch slope. Last, our study shows that, in some conditions, an excursion of 30Hz is insufficient to meet the standard JND criterion, which is 100 times greater than the estimate reported in [4]. We suggest that the fixed and arbitrary threshold of 75% for calculating JND should be revisited, as it does not account for chance performance in tasks with more than two response options.

6. Conclusions

The present study shows that pitch perception in speech is significantly more difficult than in its nonspeech analogs, such as pulse trains and pure tones. We also found that perceptibility of pitch contours is modulated by their direction, the area of the pitch range they occur in, and even segmental content. As a whole, these results motivate a reconsideration of findings from prior studies as appropriate models of human pitch perception in naturalistic speech. Furthermore, these results inform future research and design choices in human speech perception, speech processing, speech synthesis, and human-computer interaction.

7. References

[1] Bates, Douglas, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

[2] Bent, T., Bradlow, A. R., and Wright, B. A. (2006). The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 32(1), 97–103. https://doi.org/10.1037/0096-1523.32.1.97

[3] Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

[4] Flanagan, J. L., and Saslow, M. G. (1958). Pitch Discrimination for Synthetic Vowels. *The Journal of the Acoustical Society of America*, 30(5), 435–442. https://doi.org/10.1121/1.1909640

[5] Jongman, A., Qin, Z., Zhang, J., and Sereno, J. A. (2017). Just noticeable differences for pitch direction, height, and slope for Mandarin and English listeners. *The Journal of the Acoustical Society of America*, 142(2), EL163–EL169. https://doi.org/10.1121/1.4995526

[6] Klatt, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *The Journal of the Acoustical Society of America*, 53(1), 8–16. https://doi.org/10.1121/1.1913333

[7] Liu, C. (2013). Just noticeable difference of tone pitch contour change for English- and Chinese-native listeners. *The Journal of the Acoustical Society of America*, 134(4), 3011–3020. https://doi.org/10.1121/1.4820887

[8] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.