9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Exploring prosodic and conversational context factors in pitch perception

Margaret Zellers[1], *Antje Schweitzer*[2]

[1]Institut für Skandinavistik, Frisistik, und Allgemeine Sprachwissenschaft,
Christian-Albrechts-Universität zu Kiel, Germany
[2]Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany

mzellers@isfas.uni-kiel.de, antje.schweitzer@ims.uni-stuttgart.de

Abstract

Listeners use pitch information to contextualize and interpret what they hear in conversation, but contextualization requires a frame of reference in terms of both acoustic information and conversational structure. We investigate how different acoustic features such as fundamental frequency (F0), intensity, and duration, as well as different contexts for listening to speech (i.e. in isolation versus adjacent to another conversational turn) relate to listeners' perception of pitch in speech. Following a perception experiment in which listeners gave pitch ratings for individual turns or pairs of turns drawn from a corpus, we explore the relationship of prosodic features to listeners' judgments. While whole-turn F0 appears to be most relevant to judgment of turns in isolation, pitch and intensity in the region of the transition are prioritized in the turn-comparative judgments.

Index Terms: pitch, perception, conversation

1. Introduction

Listeners in conversation use pitch information in order to contextualize what they hear; that is, to infer meaning beyond lexical content and make "situated interpretations" [1]. Such interpretations take into account the form as well as the sequential position of a turn, both of which are important for understanding which conversational action a speaker is taking [2, 3]. Pitch can contribute to communicating, for example, solidarity among speakers [4], agreement or disagreement [5], or the possibility of problematic disjunction between speakers' goals [6]. [7] report that the function of turns with question structure can be identified by whether they have high or low initial pitch (i.e. not near the speaker's median pitch, and thus mismatching the context). [8] show that pitch contours in backchannels tend to be highly similar to the pitch contour at the end of the turn they follow, thus contributing to cohesion in conversational interaction. A growing body of evidence suggests that speakers accommodate or converge prosodically to one another during some phases of discourse, although the underlying mechanisms and functions of such convergence are not yet clear [9, 10, 11, 12].

Thus listeners must be able to integrate pitch features as part of the speech perception process. However, the reference frame for their integration is less clear. It seems plausible that listeners perceive pitch relative to a speaker's overall pitch: they can reliably identify the location of fundamental frequency (F0) values relative to an individual speaker's range [13], thus *normalizing* for a given speaker [14]. Taking a slightly different perspective, researchers studying pitch in conversational interaction have argued that turn pitch must be judged relative to what has come before in interaction, that is, the full sequential context; [14] calls this an *initializing* approach.

Earlier work by [15] found evidence both for a normalizing approach (in the prosodic realization of question-statement pairs) and for an initializing approach (in modeling speakers' attitudes). This work was based on automatic classifications of turn pitch using median F0 values. However, psychoacoustic studies investigating the perception of pitch demonstrate that a complex interaction arises between F0 and other acoustic characteristics of the signal [16, 17]. Since speech involves complex constellations of many acoustic features, we expect that listeners' perception of the pitch of conversational turns will be influenced by more than a simplified F0 measurement. Additionally, it is clear that top-down information modulates the processing of acoustic information at all levels of speech perception [18, 19]. Thus, while an F0-based classification is a helpful starting point, it is important to go beyond F0 to consider other prosodic and contextual features which influence listeners' interpretation of pitch in conversational turns.

The present study aims to further contribute to the discussion surrounding pitch perception in conversation by investigating how listeners interpret the pitch of spoken turns presented both in isolation and within their conversational context. Specifically, we use conditional inference trees to explore the relationship between a wider range of prosodic features, including F0, intensity, and duration, as well as variability measures, and listeners' classifications of turn pitch.

2. Data and methodology

2.1. Conversational turns

The data used for this study come from the GECO corpus [20, 21], which is comprised of 46 spontaneous conversations of approximately 25 minutes each, between previously unacquainted female German speakers. All speakers were involved in multiple dialogues. In total the corpus consists of approximately 21 hours of conversational speech on subjects of the speakers' choice.

From this corpus, [15] selected turn transitions, i.e. cases where one speaker took the turn after the other speaker. Specifically, they selected all turn transitions where the speakers' speech did not exhibit pauses longer than 0.3 s, with a minimum length of 1.5 s and no more than 0.5 s overlap with the other speaker, and no laughter, leaving 1542 turn transitions. Subsequently, turn transitions were automatically classified by register changes (low-low, low-high, ..., to high-high), either classifying each turn relative to the respective speaker's range (normalizing approach) or classifying both turns relative to the first speaker's range (initializing approach).

2.2. Perception experiment

Out of the 1542 turn pairs investigated by [15] using automatic classification methods, 90 turn pairs were selected for use in a perception experiment. The aim of the experiment was to in-

10.21437/SpeechProsody.2018-47

vestigate which prosodic features listeners rely on when making judgments about register, and to compare listener judgments with the two automatic classification methods in order to evaluate the classifications. Thus, the turn pairs were balanced (based on the initializing classification) in terms of whether they were classified as having high, mid, or low pitch, and the pairs were balanced for which of 9 classes of turn transitions they belonged to.

The experiment consisted of two portions. In the first portion of the experiment, listeners heard pairs of turns and were asked to decide whether the second turn had higher, lower, or approximately the same pitch as the first turn, i.e. to rate the relationship of the turns across the transition. In the second half of the experiment, listeners heard individual turns in isolation, i.e. only the first or only the second speaker from such a transition. In this case they were asked to respond whether the overall pitch of the turn was high, mid, or low. The 90 turn pairs were divided into two sets of 45, such that half of the participants heard set A in their paired context and set B as individual turns, while the other half of the participants heard set B in their paired context and set A as individual turns. The order of the two tasks was kept constant, since we did not want to bias listeners in the pairs task towards a speaker-normalizing judgment through the influence of voices heard in the classification task for the individual turns. In some cases, overlapping material such as backchannels were removed from the stimuli in order to ensure that each speaker could be clearly heard, reducing the duration of the turns from what was originally calculated in the automatic selection process; the individual turns used in the experiment thus ranged in duration from 1.06-8.77 seconds. Since the conversations were originally recorded in separate channels, for the experiment, the turn pairs were merged into a single channel. All stimuli were also normalized for intensity; paired stimuli were normalized together, thus retaining relative intensity differences.

The experiment was implemented using Praat's Experiment MFC [22], and carried out in a sound-attenuated booth at the University of Stuttgart. Twenty monolingual native speakers of German (17 female) participated; none reported any hearing or language impairments.

2.3. Prosodic measurements

In order to evaluate the effect that different prosodic features had on listener judgments, we took a number of prosodic measurements for each turn using Praat. The measurements are listed in Table 1. A number of global pitch and intensity characteristics were measured for each full turn. Additionally, these characteristics were measured in a "short" segment of the turn: the last 1.06 seconds (for first pair parts) or first 1.06 seconds (for second pair parts) of the turn. The duration of 1.06 seconds was chosen since it is the duration of the shortest individual turn in the experiment; the "short" measurements are thus intended as a normalization of the turns on the basis of duration.

Pitch values were measured in semitones with a reference value of 1 Hz. Since we are interested in the degree to which listeners normalize to a speaker's own range, we include both the directly measured value, as well as a value normalized to the speaker's median pitch, as calculated in [15]. The intensity values had already been normalized across all stimuli when the perception experiment was constructed. In addition to the automatic measurements, the first author, a trained phonetician, listened to all of the turns and determined whether the end of turn, defined as the last 3 syllables, involved audible creak or glottalization that could not be attributed to segmental structure.

Parameter	**Description**
In individual data	
median pitch	median pitch value (semitones from 1 semitone) (full and short)
pitch sd	standard deviation of pitch values (full and short)
corr median pitch	median pitch value with speaker median pitch subtracted (full and short)
mean intensity	mean intensity of turn (decibels) (full and short)
intensity sd	standard deviation of intensity values (full and short)
duration	duration of turn (seconds)
creak	auditorily perceptible creak/glottalization at end of turn (binary Y/N)
In pairs data	
pitch diff	median pitch of second turn minus median pitch of first turn (full and short)
pitch sd diff	pitch sd of second turn minus pitch sd of first turn (full and short)
corr pitch diff	calculated like pitch diff, but with speaker-corrected values (full and short)
intensity diff	intensity of second turn minus intensity of first turn (full and short)
intensity sd diff	intensity sd of second turn minus intensity sd of first turn (full and short)
duration diff	duration of second turn minus duration of first turn

Table 1: *Description of prosodic measurements. For the pitch and intensity measurements, two sets of measurements were taken: one for the full turn, and one for the "short" region as described in section 2.3.*

2.4. Inference tree calculation

In order to explore the relationship of the prosodic features of the turns to listeners' responses, conditional inference trees were calculated using the R statistical platform [23] and package "party" [24]. The trees were trained to predict listeners' ratings using the prosodic features in Table 1 as predictors, once for the individual ratings, and once for the transition ratings. The input for the individual tree additionally included whether the turn was a first pair part or a second pair part (since this influenced the location of the short measurements), and both trees included coding for whether the turns were questions or statements (as derived from the GECO transcriptions). The function `ctree()` calculated binary-splitting classification trees by identifying which variable was statistically most closely associated with the listener ratings, and then making a binary split in the data; the process was repeated recursively until no split could be made that reached the threshold significance level α = .05. Thus the hierarchy of features chosen in the tree reflects the potential importance of each feature for listeners' judgments.

The final tree also indicates the distribution of responses left at each bottom node, i.e. it can easily be seen in the tree which ratings are typical of each feature combination.

3. Results

3.1. Individual judgments

The decision tree for the prosodic features associated with listener judgments of individual turn pitch is shown in Figure 1. The feature which is best associated with listeners' judgments, comprising the three top-level splits, is the speaker-corrected median pitch for the turn. For turns with higher pitch overall, duration also played a role in distinguishing turns which were rated as having high pitch (node 8); longer turns were more likely to be rated high (node 10) than shorter turns (node 9). For turns with overall lower pitch, mean intensity played a role (node 3); turns with higher intensity were more likely to be rated as having low pitch (node 5), while those with lower intensity were mostly rated as medium pitch (node 4), although the split in this case is not very well-defined (in that the difference in distribution between responses falling under these nodes is not very large).

3.2. Paired turn judgments

The decision tree for the prosodic features associated with listener judgments of pitch in turn pairs is shown in Figure 2. In this tree, we see the effect of the "short" speaker-corrected median pitch (nodes 1 & 2), which is more influential than the full-turn pitch values, although the full-turn values still come into the model (nodes 3 & 11). Additionally, intensity variability in the "short" region contributes to judgments in cases where there is relatively little difference in median pitch in this region between turns (node 6). In these cases, higher variability in intensity in the second turn (i.e., larger standard deviation in intensity) as compared to the first turn is associated with the second turn being rated as having lower pitch (node 10).

4. Discussion

In both the individual and paired turns, pitch values adjusted by the speaker's median were closely associated in the trees with how listeners assigned their pitch ratings. However, the relationships between features were different in the individual versus the paired cases, suggesting that listeners used somewhat different strategies in the two tasks.

4.1. Interaction of acoustic features

Mean intensity (in the individual judgments) and relative standard deviation of intensity (in the pairs judgments) made significant contributions to the classification trees. In the individual tree (Fig. 1), higher mean intensity with a pitch near or below the speaker's median was associated with more judgments of low pitch. This may mean that higher intensity allowed listeners to make more accurate judgments of pitch. In the pairs tree (Fig. 2), however, it is greater *variability* of intensity in the ("short" region of the) second turn relative to the first that is associated with judgment that the second turn has lower pitch than the first. It is unclear why this should be the case. The node representing this result (node 10) is rather small, containing only 34 judgments; further analysis of these specific cases may shed light on the role of intensity variability in relative pitch ratings.

4.2. Prioritizing information

Listener judgments were strongly associated with speakers' (corrected) median pitch in both tasks. However, there appears to be a difference in the region of importance for the pitch values. In the individual judgments, the pitch values for the overall turn were the most strongly associated with listener judgments. However, in the turn pair judgments, the measurements which were limited to the "short" region performed better in the analysis. In other words, in the turn pair items, where listeners had to make a direct comparison, pitch in the vicinity of the turn boundary was more influential than an overall impression of the turn. However, the tree for the pairs judgments also included the comparative speaker-normalized pitch for the whole turns. Thus the judgments do not appear to be a matter of excluding information, but rather reflect a difference in cue weighting.

One possible confound for this explanation of the difference in perception between the individual turns and the turn pairs is that listeners might have prioritized pitch information near the ends of turns. Thus the location of the "short" measurements (which were at the beginning of the turn for the second pair parts) could have obscured their effect in the model. The inclusion of the pair part as a predictor in the individual model was designed to counteract this possible problem, but since the pair parts and median pitch were balanced according to our automatic classifications, any effect of the pair part may have been obscured by the method in which the tree is calculated. In order to check this, we calculated separate trees for the first and second pair parts to see if the "short" measurements would then have a stronger effect on either model. However, this was not the case. Additionally, creak, which was only annotated at the ends of turns, did not make a statistically significant contribution to the tree; an association of creak with "low" judgments might have been expected if listeners were especially attuned to the ends of turns. Thus, we do not find evidence that listeners prioritized pitch information at turn ends.

4.3. Relevance of normalization and initialization

In both trees, the F0 features that were prioritized by the model were those which had been normalized by speaker (i.e. *corrected median pitch* in the individual model, and *corrected pitch difference* and *short corrected pitch difference* in the pairs model). Thus we provide additional evidence, not only that listeners are able to normalize for individual speakers, but that they do so spontaneously. Since all of the speakers in the corpus were female, it seems unlikely that this is a general reflex of listeners normalizing to a prototypical voice, as [25] have argued. If this were the case, the speaker-normalized F0 values should have been less systematic than the raw F0 values. This may be accounted for by the difference in stimuli: [25] presented steady-state vowels, while the current experiment used full conversational turns, giving our listeners more opportunity to gather speaker-specific information.

While the current study does not directly analyze the relationship of listener judgments to automatic classifications based on pure normalizing or on an initializing-adjusted approach, the difference in the prioritized pitch features in the individual versus pairs judgments suggests that the usefulness of the approach depends on the context. In the pairs task, listeners apparently focused their attention specifically on the "short" region, that is, the region in the vicinity of the transition, in order to make their judgments. This is consistent with our previous finding that, while pitch matching based on speaker normalization coincides well with the linguistic contrast of question-answer versus

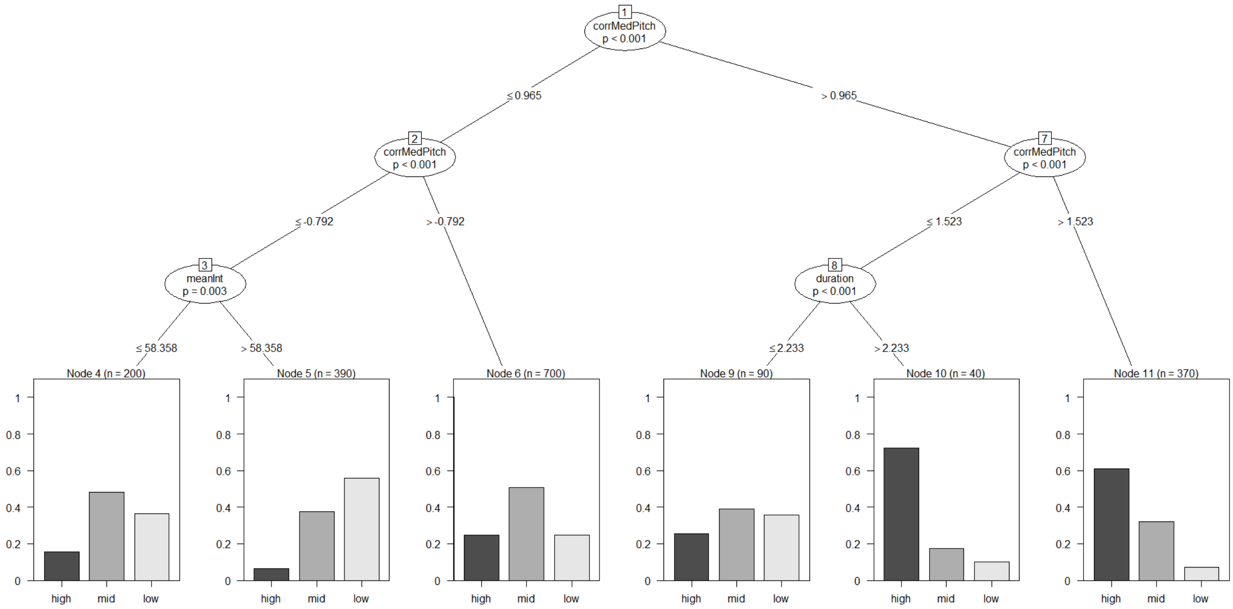

Figure 1: *Decision tree showing prosodic features associated with listener judgments of individual turn pitch.*

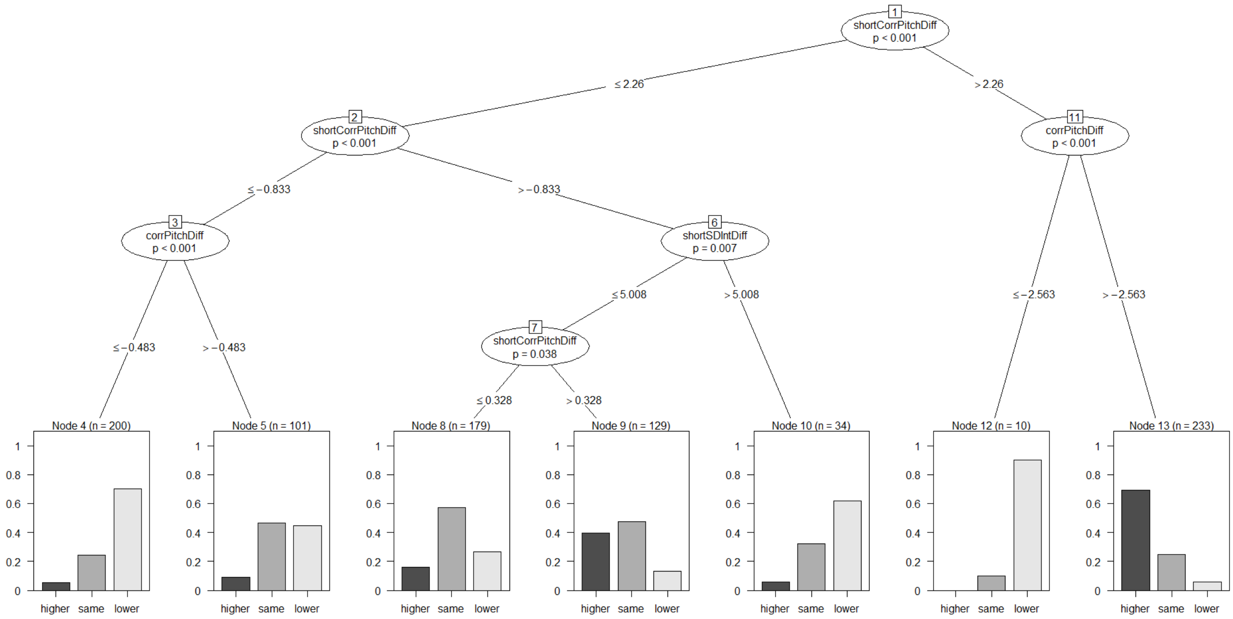

Figure 2: *Decision tree showing prosodic features associated with listener judgments of turn pitch in pairs.*

statement-statement pairs, pitch matching based on an initialization classification coincides better with the social-interactive feature of likeability [15].

5. Conclusions

We find that listeners apparently prioritize different portions of the speech signal, as well as different prosodic features, when rating pitch of individual turns compared to determining the relative pitch of turn pairs in context. This lends support for the idea that listeners adapt their pitch perception strategies depending on the context and goal of their listening. Further research is required to better define and clarify these context-specific strategies.

6. Acknowledgements

This research was supported in part by the SFB 732 *Incremental Specification in Context* (University of Stuttgart), funded by the German Research Foundation (DFG).

7. References

[1] J. J. Gumperz, "Contextualization and understanding," in *Rethinking Context: Language as an Interactive Phenomenon*, A. Duranti and C. Goodwin, Eds. Cambridge, UK: Cambridge University Press, 1992, pp. 229–252.

[2] E. A. Schegloff, "Reflections on quantification in the study of conversation," *Research on Language and Social Interaction*, vol. 26, pp. 99–128, 1993.

[3] ——, "Accounts of conduct in interaction: Interruption, overlap and turn-taking," in *Handbook of Sociological Theory*, J. H. Turner, Ed. New York: Plenum, 2002, pp. 287–321.

[4] E. Couper-Kuhlen and M. Selting, "Towards an interactional perspective on prosody and a prosodic perspective on interaction," in *Prosody in Conversation: Interactional Studies*, E. Couper-Kuhlen and M. Selting, Eds. Cambridge: Cambridge University Press, 1996, pp. 11–56.

[5] R. Ogden, "Phonetics and social action in agreements and disagreements," *Journal of Pragmatics*, vol. 38, pp. 1752–1775, 2006.

[6] M. Zellers and R. Ogden, "Exploring interactional features with prosodic patterns," *Language & Speech*, vol. 57, no. 3, pp. 285–309, 2014.

[7] M. A. Sicoli, T. Stivers, N. J. Enfield, and S. C. Levinson, "Marked initial pitch in questions signals marked communicative function," *Language & Speech*, vol. 58, pp. 204–223, 2014.

[8] J. Gorisch, B. Wells, and G. J. Brown, "Pitch contour matching and interactional alignment across turns: An acoustic investigation," *Language & Speech*, vol. 55, pp. 57–76, 2012.

[9] H. Giles, J. Coupland, and N. Coupland, *Contexts of Accommodation*. Cambridge, UK: Cambridge University Press, 1991.

[10] M. J. Pickering and S. Garrod, "The interactive-alignment model: Developments and refinements," *Behavioral and Brain Sciences*, vol. 27, 2004.

[11] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication*, vol. 58, pp. 11–34, 2014.

[12] J. Michalsky and H. Schoormann, "Pitch convergence as an effect of perceived attractiveness and likability," in *Proceedings of 18th Interspeech, Stockholm, Sweden*, 2017, pp. 2253–2256.

[13] D. N. Honorof and D. H. Whalen, "Perception of pitch location within a speaker's F0 range," *Journal of the Acoustical Society of America*, vol. 117, 2005.

[14] D. R. Ladd, *Intonational Phonology*, 2nd ed. Cambridge, UK: Cambridge University Press, 2008.

[15] M. Zellers and A. Schweitzer, "An investigation of pitch matching across adjacent turns in a corpus of spontaneous German," in *Proceedings of 18th Interspeech, Stockholm, Sweden*, 2017, pp. 2336–2340.

[16] B. C. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Bingley, UK: Brill, 2012.

[17] J. 't Hart, R. Collier, and A. Cohen, *A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody*. Cambridge, UK: Cambridge University Press, 2006.

[18] M. H. Davis and I. S. Johnsrude, "Hearing speech sounds: Top-down influences on the interface between audition and speech perception," *Hearing Research*, vol. 229, no. 1, pp. 132 – 147, 2007.

[19] E. Balaguer-Ballester, N. R. Clark, M. Coath, K. Krumbholz, and S. L. Denham, "Understanding pitch perception as a hierarchical process with top-down modulation," *PLoS Computational Biology*, vol. 5, no. 3, p. e1000301, 2009.

[20] A. Schweitzer and N. Lewandowski, "Convergence of articulation rate in spontaneous speech," in *Proceedings of 14th Interspeech, Lyon, France*, 2013, pp. 525–529.

[21] ——, "Social factors in convergence of F1 and F2 in spontaneous speech," in *Proceedings of 10th International Seminar on Speech Production, Cologne*, 2014.

[22] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer [computer program]," 2017, version Praat 6.0.24. [Online]. Available: http://www.praat.org/

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: https://www.R-project.org/

[24] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, 2006.

[25] J. Bishop and P. Keating, "Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex," *Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1100–1112, 2012.