Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

A shared control parameter for F0 and intensity

Sam Tilsen

Cornell University
tilsen@cornell.edu

Abstract

Fundamental frequency of vocal fold vibration (F0) and acoustic intensity are correlated for physiological and linguistic reasons. Previous studies have established that aerodynamic factors and vocal fold control mechanisms are partly responsible for the correlation. Intonational accents are also a source of co-variation of F0 and intensity. This paper addresses the question of whether these physiological and linguistic mechanisms are sufficient to account for observable relations between F0 and intensity in a carefully controlled context. Analyses of over 14,000 H*+L pitch accents from an imitation study indicate that in addition to physiological and linguistic mechanisms, there is a shared control parameter that induces covariation in F0 and intensity. This parameter is proposed to reflect variation in attention, cognitive effort, and/or arousal.

Index Terms: pitch accents, F0, acoustic intensity, motor control, imitation.

1. Introduction

How do speakers control phonetic parameters of speech such as fundamental frequency (F0) and acoustic intensity? Although speakers can manipulate these parameters independently, there are physiological factors (aerodynamic forces and vocal fold control mechanisms) that induce some degree of F0-intensity correlation. Furthermore, linguistic units such as intonational accents may specify both pitch targets and intensity targets, heightening correlations in analyses conducted across accent categories.

This paper presents an analysis of F0-intensity correlations calculated from over 14,000 H*+L pitch accents produced by 39 speakers in a pitch accent imitation study, addressing the question of whether previously studied physiological and linguistic mechanisms are sufficient to understand F0-intensity covariation. The results point to the existence of an additional mechanism: observed correlations were too variable across speakers to be attributable to physiological effects, and because all accents belonged to a single intonational accent category, linguistic factors cannot be responsible. Thus the results suggest that another mechanism—a shared control parameter for F0 and intensity—is needed in models of prosody generation.

1.1. Background

One source of F0-intensity correlation derives from physiological mechanisms, which can be grouped into two categories. The first category involves aerodynamic and acoustic effects. An increase in subglottal pressure increases the transglottal pressure gradient, thereby increasing vocal fold vibration rate (F0) and amplitude, resulting in higher intensity harmonics and resonances [1]–[3]. Furthermore, higher F0 results in more glottal excitation pulses per unit time, thereby increasing the root-mean-square intensity (I_{rms}). In this paper *intensity* refers to I_{rms} rather than instantaneous intensity, since I_{rms} better reflects loudness, i.e. perceived intensity. A second category of physiological mechanisms involves muscles which influence both F0 and intensity: the cricothyroid, lateral cricoarytenoid, and vocalis. Experimental studies have associated increased activity of these muscles with increases in fundamental frequency and intensity [1], [3], [4].

In contrast to physiological mechanisms, a linguistic source of F0-intensity correlation is variation in pitch accentuation. Pitch accents have been associated with increased intensity in general, and experimental results suggest that different accentual categories are associated with differences in intensity [5]–[9]. Thus analyses which are conducted across different accent categories or which incorporate measures from both accented and unaccented syllables can heighten observed F0-intensity correlations.

Variation in paralinguistic communication, particularly regarding the cognitive/emotional state of the speaker, is also a potential basis for correlation between F0 and intensity [9]–[15]. The mechanism for how cognitive states could generate F0-intensity correlations is not entirely clear. One possibility in line with the results of this study is that cognitive states influence a shared control parameter which modulates F0 and intensity.

1.2. The current study: pitch accent imitation

This paper presents a novel analysis of data collected in an intonational pitch accent imitation experiment. The experiment was designed to test the *pitch gestures hypothesis*, i.e. the hypothesis that intonational accents involve control of pitch in a manner analogous to how oral articulatory gestures are controlled [16]–[18]. On each trial, speakers heard the same synthesized name ('*Manima*) with parametrically varied H*+L accents, and they imitated those accents while producing the name in a carrier phrase. Analyses of stimulus-induced variation in F0 contours reported in [19] supported the pitch gestures hypothesis. The key findings are summarized in Figure 1: only F0 peak and F0 end manipulations in the stimuli had significant effects on response parameters; other stimulus manipulations (not shown) had less substantial effects.

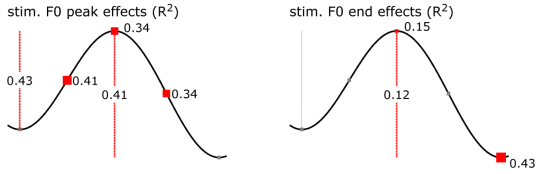

Figure 1: *Response variance explained by F0 peak and F0 end of stimulus accents.* R^2 are shown from left to right for the following response parameters: F0 rise, F0 rise speed, F0 fall, F0 peak, and F0 fall speed. See [19] for more detail.

 doi: 10.21437/SpeechProsody.2016-219

The observed effects of stimulus F0 peak and F0 end are important because the pitch gestures hypothesis holds that these parameters correspond to the targets of H and L pitch gestures, respectively. Browman & Goldstein [16] originally suggested that lexical tones and pitch accents could be conceptualized as articulatory gestures, and recent studies have shown that intonational pitch accent gestures are coordinated with vocalic gestures in German and Italian [20], [21], just as consonantal gestures are. Thus there is mounting evidence that intonational accents are pitch gestures.

1.3. Hypotheses

This paper investigates whether F0-intensity correlations are due solely to physiological mechanisms and accentual category differences, or whether additional control mechanisms are involved. The imitation paradigm is particularly well suited to addressing this question because only one accentual category (H*+L) was involved and thus observed correlations cannot be attributed to co-variation across accent categories. Hence the analyses in this study test the following hypothesis:

> *Hypothesis*: a shared parameter modulates control of both F0 and intensity.
>
> *Predictions*: measures of response F0 and I_{rms} will be correlated, but the strengths of these correlations will vary substantially across speakers.

Figure 2 compares this hypothesis with two alternative scenarios. When F0 and intensity are observed in responses from varying accentual categories and phrasal contexts (Figure 2, top), a strong correlation is expected due to context- and category-specific intensity modulation.

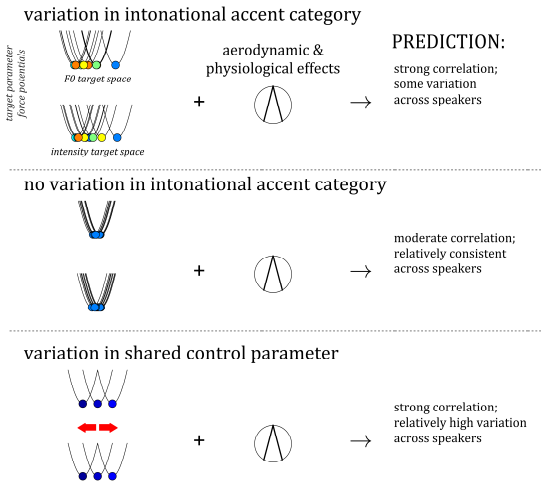

Figure 2: *Comparison of shared control hypothesis with alternative scenarios*. Top: strong correlations arise in data from multiple accent categories. Middle: physiological effects predict relatively consistent correlations across speakers within one accent category. Bottom: a shared control parameter leads to substantial interspeaker variation within one category.

However, within a single accent category, the null hypothesis holds that only physiological effects are present and correlation strengths should be relatively uniform across speakers (Figure 2, middle). In contrast, if a shared control parameter modulates F0 and intensity, a large degree of interspeaker variation in correlation strengths is predicted, even within a single accent category (Figure 2, bottom).

2. Method

2.1. Participants and Procedure

Forty-six native speakers of English participated in a one-hour session. Three participants were excluded before analyses for failing to follow instructions. Of the remaining 43 there were 18 males and 25 females; these participants imitated male- and female-voice stimuli, respectively.

Participants were seated in front of a computer screen in a sound-attenuating booth and were recorded with a head-mounted microphone. Stimuli were delivered from computer speakers using Matlab. On each trial participants heard a synthesized token of the name *Manima* and then produced the phrase *We will lay Manima near a wall*. All stimuli were presented in random order, and this was repeated 9 or 10 times over the experiment.

Participants were given two primary goals: (1) to imitate the pitch of the name *Manima* as accurately as possible, and (2) to produce the sentence without hesitating. After each trial participants received a score regarding the accuracy of their imitation and were warned if a hesitation was detected. The score was calculated as follows. The F0 contour over the phrase was obtained with normalized cross-correlation pitch tracking [22], [23], then outlying values were removed and the contour was smoothed. The absolute difference between the stimulus F0 contour and phrase contour was calculated at 5 ms steps to determine an optimal alignment, and the minimum absolute difference served as a raw imitation score. The raw scores from all preceding trials were then converted to z-scores and the most recent z-score was linearly mapped from [-2, 2] to [1, 100]. Hesitation warnings were given when a gap in voicing longer than 40 ms was detected before or after the target word.

2.2. Stimuli

Stimuli were constructed using the Mbrola diphone synthesizer [24], which allows for parametric specification of segmental duration and F0 via PSOLA. American English voices *us2* and *us1* were used to synthesize male- and female-voice stimuli, respectively.

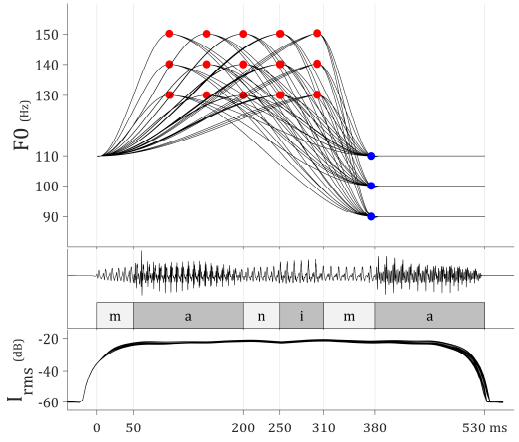

Figure 3: *Stimuli design*. Top: F0 peak, F0 peak timing, and F0 end (Hz) were varied. Middle: example waveform and segment durations. Bottom: I_{rms} (dB) contours for all 45 stimuli.

Duration parameters for the stimulus word (cf. Figure 3) were drawn from averages over several model productions of the carrier phrase with a H*+L prenuclear accent on *Manima,* produced by a male native speaker. 90 versions of the accent were synthesized over the same segmental sequence by varying three of four parameters: the starting F0 value (F0 onset), the F0 value the peak (F0 peak), the timing of the peak relative to the onset of the target word (F0 peakt), and the ending value of the F0 contour in the word (F0 offset). The contours were generated by fitting cubic smoothing splines, and model productions of the target word (*Manima*) guided selection of F0 parameters for the male voice. Female voice parameters were obtained by adding 110 Hz to all male parameters. Participants were assigned to one of two groups, in which either F0 onset or F0 offset were constant across all stimuli. Figure 3 illustrates the 45 stimulus F0 contours for the constant F0 onset group. Note that I_{rms} (40 ms window, Figure 3 bottom) is nearly identical across stimuli, with minor variations attributable to interactions between synthesized F0 contours and segments.

2.3. Data Processing and Analyses

Data from four participants were excluded because the majority of their F0 contours lacked either the rise or fall component of the accent, which precludes identification of an F0 peak. Three of these may have interpreted the stimulus as a H* accent: they delayed the F0 fall until the end of the carrier phrase. The other excluded speaker inexplicably produced F0 contours beginning with a high F0 and only falling in the target word.

F0 and root-mean-square intensity (I_{rms}) contours were extracted from each response as follows. For F0 contours, the recorded audio was high-pass filtered with a 3rd order elliptical filter having 70 and 125 Hz cutoffs for male and female speakers, respectively. The fxrapt function from the Voicebox toolbox [22], [23] in Matlab was used to extract a raw F0 contour. Allowable F0 ranges for male and female speakers were 75–250 Hz and 125–450 Hz, time frames were 11 ms and 8 ms. Each contour was further processed by removing F0 values > 4 st. dev. from the mean, interpolating gaps < 40 ms, and fitting a smoothing spline. Occasionally creaky voice in responses prevented the pitch tracking algorithm from obtaining a reliable F0 estimate, hence trials with missing frames in the target word or preceding vowel were excluded (overall 8.2% of responses). RMS intensity contours were extracted with a 40 ms window and 5 ms time step, smoothed with a 40 ms rectangular window, and transformed to a dB scale.

Segmentation was conducted through forced-alignment using the HTK-HMM toolbox [25]. For each speaker, 10 randomly selected trials were hand-segmented in Praat and used to train 5-state HMMs. A first-pass forced alignment was conducted, and then HMMs were retrained using responses with segment durations < 1.5 st. dev. from the mean. Second-pass forced alignments were obtained and guided identification of the following pitch gesture landmarks: temporal locations and values of F0 contour peak, onset minimum, and offset minimum. From these measures rise/fall duration, range, and speed (range × duration^{-1}) were calculated.

Two approaches to analysis of F0-intensity correlations are reported below: a parameter-space analysis and a principal components-space analysis. Note that "correlation" is used in two senses: as the Pearson correlation coefficient, but more often in a looser sense referring to the linear dependence of I_{rms} measures on a set of F0 measures, as estimated from linear regressions. In this latter sense R^2 values of the regression model are treated as metrics of correlation strength.

The parameter-space analysis involves the following parameters of F0 and I_{rms} contours: F0 onset, F0 peak, F0 offset, F0 rise range, F0 fall range, mean F0, I_{rms} minimum, I_{rms} maximum, I_{rms} range, and I_{rms} mean. The I_{rms} and F0 mean were calculated from only the portion of each contour within the target word *Manima*. For correlations of data pooled across speakers, parameters were z-score normalized within speakers.

The principal components-space analyses were conducted as follows. F0 and I_{rms} contours over the target word were linearly time-warped with cubic interpolation to the experiment-wide median target word duration (604 ms). For each speaker a covariance matrix was calculated from the matrix of time-normalized contours after subtracting column means, eigenvectors of the covariance matrix were sorted according their eigenvalues, and the eigenvalues were used to project the contours onto principal components-space.

3. Results

3.1. F0-I_{rms} correlations: parameter-space

Intensity (I_{rms}) and F0 parameters were correlated weakly in data pooled across speakers, but exhibited a substantial range of correlation strengths in by-speaker analyses. Figure 4 (left) shows that the highest magnitude I_{rms}-F0 correlation in the pooled data was between mean I_{rms} and mean F0 (ρ = 0.24).

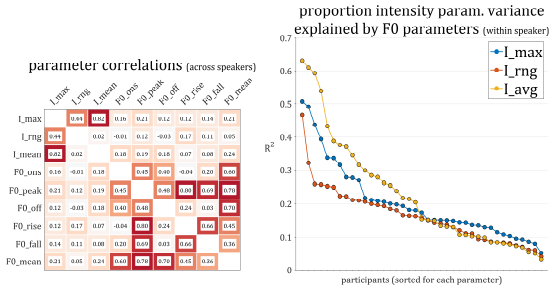

Figure 4: *Parameter-space analysis of F0-I_{rms} correlation.* Left: correlation coefficients in pooled data; color indicates magnitude. Right: R^2 from regressions of I_{rms} parameters by all F0 parameters, participants sorted by R^2 for each parameter.

Substantial cross-speaker differences were observed in the proportion of variance explained by the six F0 parameters in regressions of the I_{rms} parameters (Figure 4, right). F0 parameters did the best job of accounting for variance in mean I_{rms}, with R^2 values in the range [0.03, 0.63]. F0 parameters accounted for more than 20% of average I_{rms} variance in about half of the speakers. The wide range of correlation strengths supports the shared control parameter hypothesis: the range of correlation strengths across speakers is too large to be due to physiological mechanisms. Because all speakers employed H*+L accents, variation in accentual category cannot account for the wide range of correlations either.

3.2. F0-I_{rms} correlations: principal components space

I_{rms} and F0 principal components exhibited a substantial range of correlation strengths in by-speaker analyses, as shown in Figure 5. Notably, interspeaker differences in the cumulative proportion of contour variance explained with each principal component were larger for I_{rms} compared to F0 (Figure 5, left): for several speakers one component accounted for greater than

80% of I_{rms} contour variance, while for most speakers even the first two components accounted for less than 80%.

Substantial cross-speaker differences were observed in correlation strengths derived from regression of I_{rms} components by the first six F0 components (Figure 5, right). As in the parameter-space analysis, the wide range of variation in correlation strengths supports the shared control parameter hypothesis: a purely physiological origin for the correlation should not produce such extreme interspeaker variation.

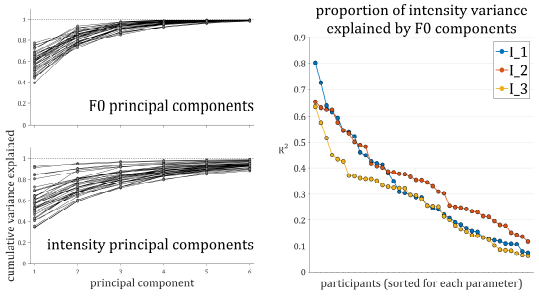

Figure 5: *Principal components-space analysis of F0-I_{rms}.* Left: cumulative variance explained by the first six principal components of F0 (top) and I_{rms} (bottom). Right: proportion of variance explained in each of the first three I_{rms} components by the first six F0 components, sorted by R^2 for each parameter.

3.3. Comparison of parameter- and pca-correlations

When correlation strengths between F0 and I_{rms} in parameter-space and pca-space are compared, the pca-space correlations are substantially stronger. Figure 6 (left) shows the maximal R^2 obtained from linear regressions of the three I_{rms} parameters with all six F0 parameters, and the maximal R^2 obtained from regressions of the first three I_{rms} principal components with the first six F0 principal components, respectively, as predictors.

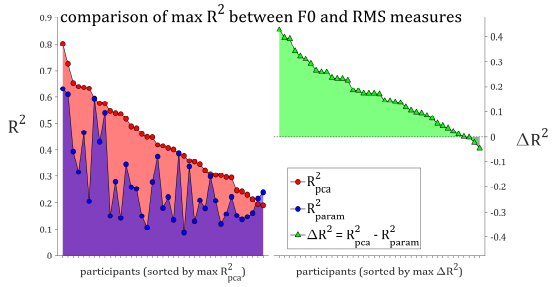

Figure 6: *Comparison of parameter- and pca-space correlation strengths.* Left: by-participant maximum R^2 values from pca and parameter space correlations, sorted by pca-space R^2. Right: difference between pca and parameter space R^2.

Note that the number of F0 principal component predictors used in the regressions matches the number of F0 predictors in parameter-space, thereby making the two analyses directly comparable. Figure 6 (right) shows the sorted differences in maximal R^2 values between the two analyses.

4. Discussion and Conclusion

Both parameter-space and pca-space analyses supported the shared control parameter hypothesis. The shared control hypothesis predicted that a wide range of correlation strengths between intensity and F0 would be observed. This was indeed the case: across speakers 5-65% of the variance in mean I_{rms} was accounted for by a set of six F0 parameters. Slightly smaller but still relatively large ranges (5-50%) were observed for the other two I_{rms} parameters, I_{rms} range and maximum. For pca-space analyses, even more variation in correlation strengths was observed across speakers: approximately 10-80% of the variance in the first three principal components of I_{rms} contours was accounted for by the first six principal components of F0 contours.

The wide range of correlation strengths is not consistent with a null hypothesis in which physiological mechanisms are the sole source of F0-I_{rms} correlation: such mechanisms predict a much more uniform effect across speakers. The other potential source of F0-I_{rms} correlation is categorical variation in intonational accent category, which could entail variation in intensity targets. Because the same pitch accent was employed throughout the entire experiment, this seems unlikely, although stimulus variation in F0 peak timing (100 to 300 ms from word onset) could have resulted in categorical differences in accent timing. To assess this possibility, supplemental analyses were conducted on two subsets of the data, with only the first and last two stimulus F0 peak timing conditions, respectively. The results of these analyses were qualitatively similar to those reported in sections 3.1-3.3, and hence categorical differences in accent timing are not a likely source of the interspeaker variation in correlation strengths.

Another alternative hypothesis worth further investigation is that speakers varied in the extent to which they perceived intensity/F0 manipulations in the stimuli, and this perceptual variation could be reflected in their imitations, inducing F0-intensity correlation. Indeed, a number of studies have shown that perception of F0 and intensity interact [26]–[29]. Because of these interactions, and because speakers may have learned associations between F0 and intensity, participants may have non-veridical perceptions of intensity variation in the stimuli. An imitation task might be conducted in tandem with perceptual tasks to assess this hypothesis in future studies.

Given the notion of a shared control parameter for F0 and intensity, one can ask how this parameter should be conceptualized, and how mechanisms associated with this parameter should be incorporated into models of prosody generation. Here it is proposed that the shared control parameter reflects unconditioned fluctuations in cognitive states, potentially associated with attentional energy (or effort, or arousal); similar fluctuations have been observed in oral articulatory timing [30] and are thus expected to modulate F0 and intensity targets as well.

Models of prosody generation in which pitch accents are viewed as gestures are advantageous from this perspective because F0 gestural targets can be dynamically biased by multiple factors, which can be conceptualized as forces. Thus speakers maintain representations of F0 and intensity targets in an articulatory control space, and the positions of those targets can be modulated by the shared control parameter, which reflects variation from a number of sources (attention, arousal, effort, etc.). Hence a prediction of this model is that speakers with greater degrees of F0-intensity correlation will exhibit more fluctuation in attention/arousal/effort and greater variability in articulatory timing.

5. Acknowledgements

I would like to thank Eric Evans, Emma Lantz, and Danielle Burgess for assistance in conducting this study.

6. References

[1] M. Hirano, J. Ohala, and W. Vennard, "The Function of Laryngeal Muscles in Regulating Fundamental Frequency and Intensity of Phonation," *J. Speech Lang. Hear. Res.*, vol. 12, no. 3, p. 616, Sep. 1969.

[2] I. R. Titze, "Regulation of vocal power and efficiency by subglottal pressure and glottal width," *Vocal Fold Physiol. Voice Prod. Mech. Funct.*, pp. 227–238, 1988.

[3] I. R. Titze, *Principles of voice production*. National Center for Voice and Speech, 2000.

[4] I. R. Titze and D. W. Martin, "Principles of voice production," *J. Acoust. Soc. Am.*, vol. 104, no. 3, pp. 1148–1148, 1998.

[5] M. E. Beckman, *Stress and non-stress accent*, vol. 7. Walter de Gruyter, 1986.

[6] D. Bolinger, "A theory of pitch accent in English," *Word*, vol. 14, no. 2–3, pp. 119–149, 1958.

[7] A. Rosenberg and J. Hirschberg, "On the correlation between energy and pitch accent in read English speech.," in *INTERSPEECH*, 2006.

[8] A. M. Sluijter and V. J. Van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *J. Acoust. Soc. Am.*, vol. 100, no. 4, pp. 2471–2485, 1996.

[9] C. Gussenhoven, *The phonology of tone and intonation*. Cambridge: Cambridge University Press, 2004.

[10] C. Pereira and C. I. Watson, "Some acoustic characteristics of emotion.," in *ICSLP*, 1998.

[11] P. N. Juslin, K. R. Scherer, J. Harrigan, R. Rosenthal, and K. R. Scherer, "Vocal expression of affect," *New Handb. Methods Nonverbal Behav. Res.*, pp. 65–135, 2005.

[12] J. Pittam and K. R. Scherer, "Vocal expression and communication of emotion," *Handb. Emot.*, pp. 185–197, 1993.

[13] K. R. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion," *Handb. Affect. Sci.*, pp. 433–456.

[14] K. R. Scherer, "Vocal affect expression: a review and a model for future research.," *Psychol. Bull.*, vol. 99, no. 2, p. 143, 1986.

[15] D. R. Ladd, *Intonational phonology*. Cambridge: Cambridge University Press, 2008.

[16] C. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, no. 2, pp. 201–251, 1989.

[17] C. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3–4, pp. 155–180, 1992.

[18] E. Saltzman and K. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecol. Psychol.*, vol. 1, no. 4, pp. 333–382, 1989.

[19] S. Tilsen, D. Burgess, and E. Lantz, "Imitation of intonational gestures: a preliminary report," *Cornell Work. Pap. Phon. Phonol. 2013*, pp. 1–17, 2013.

[20] D. Mucke, H. Nam, A. Hermes, and L. Goldstein, "Coupling of tone and constriction gestures in pitch accents," in *Consonant Clusters and Structural Complexity*, Berlin: Mouton de Gruyter, 2012, pp. 205–230.

[21] H. Niemann, D. Mücke, H. Nam, L. Goldstein, and M. Grice, "Tones as Gestures: the Case of Italian and German," *Proc. ICPhS XVII*, pp. 1486–1489, 2011.

[22] M. Brookes, "Voicebox: Speech processing toolbox for matlab," *Softw. Available Mar 2011 Www Ee Ic Ac Ukhpstaffdmbvoiceboxvoicebox Html*, 1997.

[23] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding Synth.*, vol. 495, p. 518, 1995.

[24] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken, "The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, vol. 3, pp. 1393–1396.

[25] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book," *Camb. Univ. Eng. Dep.*, vol. 3, p. 175, 2002.

[26] D. G. Kemler Nelson, "Processing integral dimensions: The whole view.," 1993.

[27] J. W. Grau and D. K. Nelson, "The distinction between integral and separable dimensions: Evidence for the integrality of pitch and loudness.," *J. Exp. Psychol. Gen.*, vol. 117, no. 4, p. 347, 1988.

[28] R. D. Melara and L. E. Marks, "Interaction among auditory dimensions: Timbre, pitch, and loudness," *Percept. Psychophys.*, vol. 48, no. 2, pp. 169–178, 1990.

[29] J. G. Neuhoff, J. Wayand, and G. Kramer, "Pitch and loudness interact in auditory displays: Can the data get lost in the map?," *J. Exp. Psychol. Appl.*, vol. 8, no. 1, p. 17, 2002.

[30] S. Tilsen, "Structured nonstationarity in articulatory timing," *Proc. 18th Int. Congr. Phon. Sci.*