MODELING PITCH RANGE VARIATION WITHIN AND ACROSS SPEAKERS: PREDICTING F_0 TARGETS WHEN "SPEAKING UP"*

Elizabeth Shriberg,[1] D. Robert Ladd,[2] Jacques Terken,[3] Andreas Stolcke[1]

[1]SRI International, Menlo Park, California, USA ([ees,stolcke]@speech.sri.com)
[2]University of Edinburgh, Edinburgh, United Kingdom (bob@ling.edinburgh.ac.uk)
[3]Institute for Perception Research, Eindhoven, The Netherlands (terken@natlab.research.philips.com)

ABSTRACT

We study F0 variation produced by "speaking up", as part of a larger study of pitch range variation within and across speakers [1]. We provide a function to predict target F0 values in this "raised" mode from F0 values at corresponding locations in speech produced in a neutral mode. Targets were F0 measurements at points of low internal variability in read Dutch sentences produced by 15 speakers. Results showed that the speaker dependent variability was well described by an additive-multiplicative model in the linear frequency domain. Furthermore, across speakers, the additive and multiplicative parameters were negatively correlated. One free speaker-dependent parameter could therefore be eliminated by adding a single speaker-independent linear constraint.

1. INTRODUCTION

Pitch contours are the result of both linguistic and nonlinguistic influences. On the linguistic side, they provide the realization of particular tonal sequences; on the nonlinguistic side, they reflect the contribution of speaker characteristics (most notably speaker sex) and factors that bring about within-speaker variation (such as conversational setting, or emotional state). To extract information relevant to the realization of tonal sequences, we must be able to separate out the contribution of nonlinguistic factors to phonetic observations. This requires explicit models to relate nonlinguistic factors to pitch range variation within and across speakers.

Previous work on pitch range variation comes from a number of fields. Analyses have involved production and perception studies on intonational, tonal, and prominence-related invariances across pitch range, using different theoretical frameworks [among others, 2, 3, 4], and several models for pitch range have been proposed [5, 6, 7]. In addition, studies have described physiological correlates, overall acoustic changes, and changes associated with shouting, emotions, or stress level. Much is not yet understood, however, about how intonational, tonal, and prominence relationships are preserved across pitch range, and how these relationships vary for different speakers.

The present study is part of a large-scale investigation of pitch range variation within and across speakers [1]. We focus here on the pitch range variation produced when speakers deliberately raise their voices in the context of communicating over a (simulated) noisy telephone channel. We examine data from a relatively large number of speakers, and representing a variety of linguistic patterns. We can therefore investigate not only appropriate functions for describing pitch range relationships within a particular speaker, but also ask how well different functions can describe variation across speakers.

Specifically, we seek a "raising function" by which to relate F0 targets in the raised mode to corresponding targets in the same sentences spoken in the normal mode. We assume that a considerable portion of the pitch change obtains because speakers deliberately raise their pitch; this is the phenomenon we intend to describe. We recognize, however, that some portion of the changes will also occur as a by-product of an increase in vocal effort (the "Lombard reflex" [8]).

2. METHOD

Speech data consisted of recordings from 15 adult native speakers of Standard Dutch (7 male, 8 female). The database is described in further detail in [1]. Speakers produced multiple repetitions of various sentence types, first in a quiet room, and then with loud noise presented over headphones, so as to simulate talking over a bad overseas telephone connection. The sentence types were designed to elicit specific intonation patterns expected to have consistent and identifiable peaks and valleys at well-defined locations. They included statements of varying lengths, and statements with explicit contrasts. Examples of the five sentence types used in the current work are provided below. Accented words are shown in uppercase letters.

1. Je moet de MOOIE ROZEN in een GELE VAAS doen.
   (You should put the pretty roses in a yellow vase.)
2. Je moet de MOOIE GELE ROZEN in een VAAS doen.
   (You should put the pretty yellow roses in a vase.)
3. We zouden wel eens naar [STAD] kunnen gaan.
   (We really ought to be able to go to [city] sometime.)
4. Zij moet [DAG] gaan.
   (She has to go on [DAY].)
5. Ik zei niet [X] maar [Y].
   (I didn't say [X], but [Y].)

*This paper replaces an incorrect version, "Modeling Intra-Speaker Pitch Range Variation..." by Shriberg, Ladd, & Terken which appeared in the original hardcopy and CDROM proceedings of ICSLP 1996.

F0 values were extracted at hand-marked locations corresponding to accent peaks, half-accent peaks, valleys, and sentence-final lows.

3. MODELING

3.1. Targets Modeled

To uncover invariances across pitch range, it is optimal to use targets that show low internal variability. Consistent with previous studies [2, 7, 9], we found that all target types showed relatively invariant F0 values. That is, repeated versions of the same sentence type by the same speaker were produced using similar F0 values. Results for one speaker for sentence type 1 are shown in Figure 1.

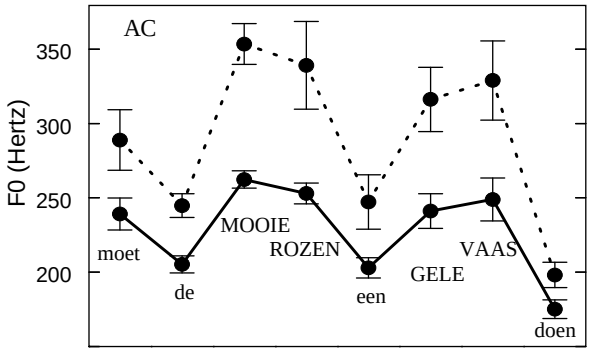

Figure 1: Mean and standard deviations for targets in the normal and raised modes for speaker AC (female), for sentence type 1.

Next, we asked whether the relationship between normal and raised targets was affected by target type. For each speaker we examined a scatterplot in which the mean normal value was plotted against the mean raised value, for all target types and sentence types. Results for a sample speaker are shown in Figure 2. As shown, with the exception of one target type, points fell roughly on a line. The exception was the final low. For many speakers (such as AC in Figure 1) the raised value was significantly higher than the normal value (as can be deduced from the error bars in Figure 1). However, raised-mode final lows were typically lower in F0 than expected based on the trend formed by the other targets (as shown for ES in Figure 2). This suggests that the raising of final lows has a different causal explanation and should be modeled separately. We therefore excluded final lows in all further analyses.

Examination of scatterplots like those in Figure 2 across speakers revealed that there was no systematic variation by target type (after removal of final lows). Therefore, we collapsed over target type for the analyses. In addition, we examined the scatterplots for an effect of sentence type. We found no consistent effect across speakers, and thus also collapsed over sentence type.

3.2. Two-Parameter Raising Functions

Since targets fell roughly on a straight line after exclusion of final lows, the data could be modeled with the simple linear function

$$R = a\,N + b$$

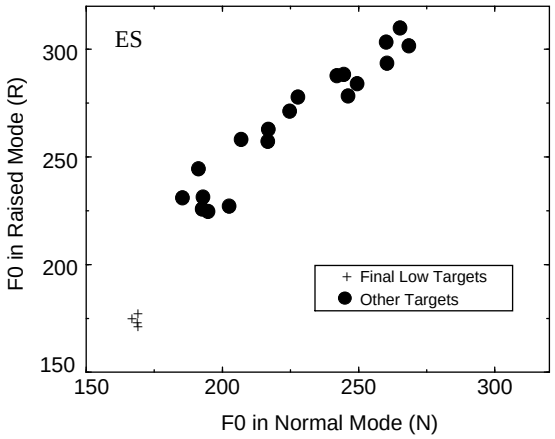

Figure 2: Normal versus raised targets for speaker ES (female) by target type.

where R is the value of a raised target, N is the value of the corresponding normal target, and *a* and *b* are free parameters.

This linear function relates normal to raised targets using two parameters in a raw F0 space. Past work, however, has proposed nonlinear functions for the scaling of F0. To investigate whether the alternative scales provided any reduction in prediction error, we compared results for the unconstrained linear function to results for two nonlinear scaling functions proposed in the literature: log and ERB. The log function corresponds to equal intervals on a ratio or semitone scale [10]. The ERB (equal rectangular bandwidth) function predicts equal intervals based on a psychophysical scale [3], and is between linear and log for the range of F0 values observed. We compared the linear function above with two corresponding functions in which R and N were replaced by their scaled values.

Results showed that none of the nonlinear functions systematically reduced prediction error over speakers, and within-speaker differences across models were quite small. Thus, there was no reason to chose a scaling function other than the identity function to describe these data. Furthermore, as explained in the next section, the linear raising function has the advantage that it can be rewritten such that its parameters have a straightforward interpretation.

3.3. Interpreting the Linear Raising Function

In the linear model, the parameter *a* is the factor (in linear space) by which a speaker's F0 range in the normal mode is expanded in the raised mode. The parameter *b* allows the F0 range to be shifted as well. However, in the standard parameterization of $aN + b$, the value of *b* is a frequency that does not correspond to an observable value; indeed it can take on impossible (e.g., negative) values. To obtain a linear raising function in which the shift parameter is interpretable, we express all frequencies as offsets to the minimum observed normal-range value (the overall minimum). After this translation of the F0 space, *b* can be interpreted as the upward F0 shift applied to the minimum normal F0.

3.4. One-Parameter Raising Functions

As outlined earlier, the two-parameter linear raising function provides a good description of the observed data. A further question is whether the data can be described using fewer than two free parameters per speaker. In general the fewer parameters a model has, the more likely it is to be explanatory as opposed to purely descriptive. In addition, when making practical use of a statistical model, one does not want unnecessary parameters because such parameters will capture noise or overfit the training data.

In light of the literature on F0 scaling, we might try to eliminate the additive (shift) component of our linear model. This would leave only a multiplicative relationship between normal and raised targets, which can be expressed as an additive model (equal distances) on a log scale. Results for this simplified model are shown in column 2 of Table 1. As a baseline, results for the two-parameter model are shown in column 1.

Speaker(sex)	R=*a*N + *b*	R=*a*N	R-R_{min}=*a*(N-N_{min})	R=*a*N+*b* (*a*, *b* tied)
AB (f)	14.86	16.72	16.01	17.73
AC (f)	10.26	17.45	10.82	18.98
ES (f)	8.12	9.15	8.19	10.60
EV (f)	8.32	8.51	8.38	8.35
IS (f)	13.93	22.05	27.67	19.86
LA (f)	24.44	25.71	24.61	29.40
LV (f)	19.54	32.54	23.92	22.32
UA (f)	11.59	12.82	11.82	12.32
Fem. ave.	**13.88**	**18.12**	**16.43**	**17.45**
JR (m)	5.90	7.12	6.98	7.82
MH (m)	15.65	15.91	16.00	15.79
RE (m)	9.71	10.63	10.06	10.29
RH (m)	16.82	16.83	17.79	16.89
RL (m)	14.49	15.57	14.90	16.00
RS (m)	9.76	10.32	10.05	10.07
RW (m)	10.51	15.19	11.79	14.48
Male Ave.	**11.83**	**13.08**	**12.51**	**13.05**

Table 1: Standard deviations for different models by speaker, with overall averages for speakers grouped by sex.

As expected, there is some loss in predictive power when the shift factor is eliminated. Notably however, this simplification affects speakers to varying degrees, depending on the magnitude of their naturally-occurring shift (see for example speakers IS and LV). Thus, eliminating the shift parameter is undesirable because it removes the models' ability to account for one of the distinct dimensions along which speakers differ.

A more principled way to remove the additive component in the linear model is to subtract the minimum value in each speaking mode (N_{min} and R_{min}). As shown in column 3 of table 1, after these two translations of the frequency space, results are much closer to those from the two-parameter linear fit on the raw data. This approach is only parsimonious, however, if N_{min} and R_{min} for each speaker can be set independently of the data to be modeled. In the present analyses, we estimated N_{min} and R_{min} from the data to be modeled (for lack of another source for the estimates). Therefore, these results should only be compared to those for the two-parameter linear model.

3.5. Raising Parameters by Speaker

So far all of our modeling efforts have involved using parameters that are specific to a particular speaker. Another approach to reducing the overall number of parameters for a set of speakers is to constrain the speaker-specific parameters by cross-speaker relationships. For example, we could try to find a universal relationship across speakers that predicts one of the raising parameters from the other. This would reduce the number of speaker-dependent parameters (to one per speaker), while adding only the fixed number of parameters necessary to describe the universal relationship.

To discern whether there was a systematic relationship over speakers, we plotted *a* against *b* obtained from the two parameter linear fits for each speaker. This plot, shown in Figure 3, suggests a negative, roughly linear relationship between *a* and *b.* Such a relationship is not merely an empirical finding, but can be explained by a universal constraint on the production of the raw data. We can assume that due to physical constraints, all F0 values across speakers have to lie within a certain range. In other words, the R/N plots for all speakers (exemplified by Figure 2), have to lie within the same rectangular area. It follows from geometry that a speaker's raising function cannot have both a high *a* (slope) and a high *b* (y-intercept) value. Rather, speakers have to trade off high F0 range expansions against low F0 shifts, and vice-versa. Furthermore, the same geometrical reasoning shows that the further away from 0 the allowed F0 range, the stronger the negative relationship between *a* and *b* will be. This is reflected in the Figure 3 by the fact that *a*/*b* points for females lie on a steeper negative slope that those for males.

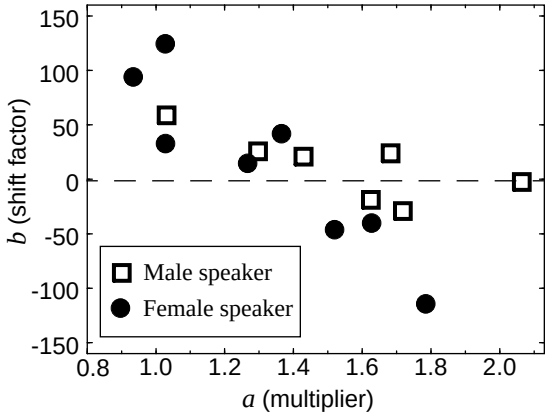

Figure 3: Values of a and *b* by speaker, for 2-parameter linear fit.

Based on this reasoning, we imposed a linear constraint relating each speaker's raising parameters by the function

$$b = l * a + m ,$$

where *l* and *m* are now speaker-independent parameters. Following the argument above, we allowed separate *l* and *m* values for males and females.

To find suitable *l* and *m* values we cannot simply do a regression in *a*/*b* parameter space, since distances in that space are not monotonically related to the prediction errors in frequency space. Instead, we chose the speaker-independent, gender-dependent *l* and *m* values such that fitting all speakers' raising functions subject to the constraint $b = l * a + m$ minimized the overall prediction error. We will refer to the model thus obtained as the "tied" model, since our speaker-independent constraint ties the *a*/*b* parameters for each speaker, both to each other and to the parameters of other speakers.

The total number of parameters in the tied model is intermediate between those of the unconstrained and the all-multiplicative models. Correspondingly, its overall prediction error is guaranteed to lie between that of the other two models, as shown in Table 1, column 4. One can also see that the overall improvement over the multiplicative model, which constrains *b* to be 0, is appreciable for female speakers, but small for males. This is to be expected from Figure 3, as the optimal *b* values for males are close to 0. Note that the results by speaker for the tied model can be better or worse than the multiplicative-only model. Since the global prediction error is minimized, the tied model avoids results by speaker that are drastically worse than those for the unconstrained linear fit.

We should mention that the simple linear constraint assumed is probably a simplification of the actual constraints on F0 raising across speakers. The relationship between *a* and *b* may not be exactly linear, or factors other than F0 range may contribute to the relationship. The general point is that it is possible to integrate speaker-dependent and speaker-independent modeling to obtain models that are more parsimonious and explanatory overall.

4. SUMMARY AND CONCLUSION

We found that speakers show high internal consistency in the production of F0 targets in both normal and raised speaking modes, and that the relationship between corresponding normal and raised targets was well described by a linear function with two free parameters. The multiplicative parameter in this function corresponds to the change in F0 excursions from normal to raised mode; the additive component corresponds to the upward shift from the bottom of the normal range to that of the raised range.

We further investigated whether a more parsimonious model could describe the data, by exploring models with only one free parameter per speaker. In one case we eliminated the shift parameter, thereby reducing the speaker-dependent model to a purely multiplicative one. As expected, this resulted in a considerable increase in prediction error for a number of speakers. As an alternative we searched for a universal relationship between the two speaker-dependent parameters, and found one such constraint. The multiplicative and additive parameters for each speaker were negatively correlated—a result which can be explained by global bounds on F0 ranges for male and female speakers.

By modeling this negative correlation as an additional, speaker-independent linear constraint, we were able to effectively eliminate one free parameter per speaker. The resulting, tied model produced predictions that were overall better than those produced by the all-multiplicative model, which also had only one free parameter per speaker. Although the one-parameter tied model cannot match results for the two-parameter linear model, the tied model is more attractive from a theoretical perspective, since it directly reflects similarities as well as differences across speakers. In addition, the tied model may be preferable from an applied perspective, since it reduces the overall number of parameters to be estimated.

5. REFERENCES

1. Ladd, D.R. and Terken, J. "Modeling inter- and intra-speaker pitch range variation," *Proc. International Congress of Phonetic Sciences*, Stockholm, 1995.
2. Liberman, M. and Pierrehumbert, J. "Intonational invariance under changes in pitch range and length," In *Language Sound Structure*, M. Aronoff and R. Oehrle (eds), Cambridge, MA: MIT Press, 1984.
3. Hermes, D. and Van Gestel, J. "The frequency scale of speech intonation," *Journal of the Acoustical Society of America*, *90*(1), 97-102, 1991.
4. Rose, P. "Considerations in the normalisation of the fundamental frequency of linguistic tone," S*peech Communication 6*, 343-351, 1987.
5. Ladd, D.R. "Metrical representation of pitch register," *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, J. Kingston and M. Beckman (eds), Cambridge: Cambridge University Press, 35-70, 1990.
6. Liberman, M., Schultz, J., Hong, S. and Okeke, V. "The phonetics of Igbo tone," *Proc. ICSLP*, 743-746, 1992.
7. Van Den Berg, R., Gussenhoven, C. and Rietveld, A. "Downstep in Dutch: implications for a model," *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, J. Kingston and M. Beckman (eds), Cambridge: Cambridge University Press, 335-359, 1990.
8. Lombard, E. "Le signe de l'elevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37, 101-119, 1911.
9. Bruce, G. *Swedish word accents in sentence perspective*, Lund: CWK Gleerup, 1977.
10. ´t Hart, J., Collier, R., and Cohen, A. *A Perceptual Study of Intonation*. Cambridge: Cambridge University Press, 1990.

ACKNOWLEDGMENTS

We thank Sjoerd Zinnemers for data annotation, Ercan Gigi for assistance with waveform analysis software, and Remco Teunen for assistance with the regression analyses. The second author was supported by a van Houten fellowship at IPO, funded by Philips. The first author was supported by an NSF-NATO postdoctoral fellowship at IPO, and by DARPA and NSF under NSF Grants IRI-9314967 and IRI-8905249. The views herein are those of the authors and should not be interpreted as representing the official policies of DARPA or the National Science Foundation.