# REPRESENTATION OF PROSODIC STRUCTURE IN SPEECH USING NONLINEAR METHODS

*Rashid Ansari** *Yi Dai** *Jian Lou** *David McNeill*** *Francis Quek**

* Dept. of EECS (M/C 154), University of Illinois at Chicago, Chicago, IL 60607, USA
** Department of Psychology, University of Chicago, Chicago IL 60637, USA

## ABSTRACT

As part of an effort to discover and quantify cues in the modalities of speech, gesture, and gaze, we are developing automated methods of extracting the primitives from the raw multimodal data, and representing and organizing them in new and flexible ways. An in-depth study of the acoustic-prosodic properties of speech such as pitch, amplitude, duration, and speaking rate, and their relation to other modalities in conveying discourse-level meaning requires the processing of a large amount of the recorded audio signal. In this paper we describe a procedure to process the fundamental frequency (F0) trace in order to (i) obtain an improved estimate, (ii) segment the processed F0 data, and (iii) find a parametric representation suitable for the multimodal analysis. The goal is to use flexible parametric representations whose parameters can be examined for correlation with suitably extracted gesture and gaze features. Examples of application of the procedure for preprocessing, segmentation and representation of the F0 samples are described.

## 1. INTRODUCTION

In this paper we describe a speech processing procedure that will be incorporated in software for automated discourse structure analysis based on gesture and speech prosody cues. It has been observed that human intonational and gestural behaviors usually function as integrated semiotic systems [11]. Analysis of gesture and speech in discourse management and investigation of correlation between the two modalities will require extraction, processing, and study of an immense amount of data. In our effort to make the task of analysis easier and less time-consuming, we are developing speech processing and computer vision tools to perform automatic feature extraction, parametric representation, and event detection that will aid the research in investigating prosody and gesture nexus.

Numerous findings on intonational characteristics of spoken discourse have been have been reported [2, 3, 4, 8, 9]. The capture and representation of prosodic information in speech is important in applications such as speech synthesis, speech compression and multi-modal discourse analysis. In speech synthesis [1, 5, 10, 13, 15, 16] the quality of synthesized speech can be improved by exploiting the ability of intonation to reliably convey the linguistic structure at the discourse level. This is accomplished by producing contextually appropriate intonational variation.

An example of an application that requires compact representation of F0 and amplitude information is a compression procedure based on lossy transform coding described in [6]. The compressed information is used to recreate sound by modifying the pitch of speech units selected from an inventory of recordings with fixed pitch contours. If one disregards speaker characteristics and speech prosody, it has been observed that the fundamental information transmission rate for a human reading text is on the order of 100 bits/sec [14]. This observation taken together with compact coding of the prosodic information can be used to synthesize speech, where the overall information is represented at significantly reduced bit rates compared with conventional techniques. The approach relies on the use of effective methods of concatenative synthesis, which consists of selecting a set of basic acoustic units, recording them in natural voice, and generating utterances by concatenating appropriately modified segments. This method relies on the ability to represent F0 and amplitude variations in a compact way using suitable basis functions. The F0 and amplitude values, computed with X-waves software, are represented over isolated voiced segments with variants of lossy transform coding. Discrete cosine and sine transforms are used to represent F0 and amplitude respectively.

We are currently exploring linkages between speech, gesture, gaze, and discourse. One area of focus is the investigation of the relationship between gesture and prosody, which remains largely unexplored. The work reported here addresses related issues in F0 representation using nonlinear methods. Structure in the F0 trace is a critical element in this investigation. Variations in F0, F0 trends at low resolution, semantically appropriate F0 contours, and location of F0 extrema are some of the properties that are potentially valuable in this analysis. One of the goals in this work is to help build a flexible signal processing tool that can be modified to vary models for representing F0 and other prosodic information.

In the next section we describe the acquisition of the audio and video data and the extraction of F0. The F0 estimates are often erroneous and we describe a method of preprocessing the F0 trace before using it in analysis. In section 3 methods for segmenting the data and determining parametric representations are described. Examples of application of the processing are presented in sections 2 and 3. The objective of these methods is to render the information into forms that can be conveniently analyzed.

## 2. EXPERIMENTAL SET-UP AND F0 PREPROCESSING

In this section we describe the experimental set-up for capturing the speech and gesture data, and the extraction of F0.

### 2.1. Experimental Set-up

In one of the experiments for gesture and speech elicitation, subjects are asked to describe their house to an interlocutor. The conversation is recorded on a Hi-8 tape using a regular camcorder. Two sets of analyses are performed on the video and audio data. In the first set of analyses speech and video processing software is used to obtain (i) the speech fundamental frequency and amplitude (in terms of the RMS value of the audio signal) from the audio signal, and (ii) the motion traces of both of the subject's hands from the video signal. The second set of analyses consists of expert transcription of the speech and gesture data. This transcription is done by expert psycholinguists, and it serves to guide the algorithm development using the cues accessible in the gestural and audio data.

### 2.2. Preprocessing the F0 Trace

The raw speech signal considered here pertains to the description of a house by a female subject. The signal is resampled at 8 KHz. We use xwaves+ software to get a preliminary estimate of the fundamental frequency (F0) and amplitude (RMS) of the speech. The RMS trace is refined with a median smoother and used in preprocessing the F0 trace.

The presence of noise in the speech recording leads to incorrect F0 estimates, due to which feature extraction can be erroneous and turn out to be misleading in the analysis. The F0 preprocessing is intended to correct the F0 values before they are used in analysis. On examining the preliminary estimate of F0 obtained directly from xwaves+ software, several F0 values are found to be incorrectly estimated, often in bursts. The incorrectly estimated F0 values in the trace considered here are found to be lower than the correct F0, usually by a factor of 2 or 3. For the female voice recorded in the experiment, the median F0 is first estimated and found to be close to 200. The incorrect F0 estimates usually assume a burst of values close to either 100 or 70.

The errors in F0 values obtained from xwaves+ can be classified into two categories: (i) False F0s, where the presence of background interference or noise causes the xwaves+ software to detect false F0s in unvoiced speech. In this case the original F0 value is often in the range 60 to 80. (ii) F0 sub-harmonics, where the speech is voiced but the F0 estimates are either half or a third of the correct value.

In our processing of the female voice we assume that F0 values that are lower than a threshold frequency $F_t = 150$ are candidates for correction. We attempt to correct these F0s, while leaving other values unchanged. The decision on whether and how to correct the F0 values is based on a closer examination of the raw speech signal.

One item of information that aids the F0 preprocessing is the value of the corresponding amplitude. Based on the average amplitude and noise pedestal, we set a threshold amplitude for accepting data as voiced and with a non-zero F0 value. If the xwaves+ F0 trace contains a suspect non-zero value with corresponding amplitude that is below threshold, then the F0 is set to 0 (unvoiced/silence). This correction is performed over a contiguous voiced segment. The amplitude that is compared with the threshold is the maximum value in the median-smoothed RMS trace within a segment.

The procedure above allows us to disregard some F0s when the signal amplitude is below threshold. However the threshold cannot be set too high as this may cause loss of some genuine low-amplitude voiced segments. Therefore the F0 trace processed as above may retain some incorrect F0s in regions of noise. In order to further identify the noisy data, we examine the original sampled speech to determine the extent of signal correlation. Assume that the F0 value in some segment is $f$ which is less than $F_t$. We examine the speech signal in that segment. The pitch period of the speech waveform should be $T = 1/f$. We now compute the normalized signal correlation for a time-shift of $T/2$ and $T/3$ and use it to not only distinguish a genuine F0 from an incorrect F0 corresponding to noisy data but also to correct its value.

Another clue to incorrect F0s is the presence of abrupt changes in F0 values within a segment. If a F0 segment contains discontinuities, with F0 values below $F_t$ on one side of the discontinuity, then a different procedure is used to correct F0. When discontinuities are encountered, we use a predictor from the region of F0 values that are likely to be correct. A predictor (one step or two steps) is used to estimate the next F0 sample. We then scale the incorrect F0 up by a factor of 2 or 3, depending on which result is closer to the prediction. Because the incorrect F0 can be at the beginning, end, or middle of the segment, we need to perform both forward and backward prediction in order to make the correction.

Figure 1 shows a 10-second section of the F0 trace and the result after preprocessing. The time index is the frame number of the video data recorded at 30 frames/sec. In Figure 2 the histogram of F0 values in the original and preprocessed F0 trace is shown.

## 3. PARAMETRIC REPRESENTATION OF F0

In pitch analysis of spoken English, F0 tends to decline over course of phrases and utterance [8]. The F0 declination represent a low-resolution trend, one of the possible trend functions of interest in analysis. We now consider a framework for a flexible representation of trend functions.

To represent the F0 information in a manner appropriate for studying correlation with gesture, it is important to capture the main low-resolution trends in F0 variation. The trend (e.g. declination) can be factored out when one wishes to examine the high-resolution component of the trace which represents fine local variations. We describe a procedure in which a F0 trend function is modeled with a small number of parameters. The trend function can be chosen as a smooth function that captures a general trend of F0 movement. One can use a linear or an exponential function to model the trend. Other basis functions such as the discrete cosine and sine functions may also be used to

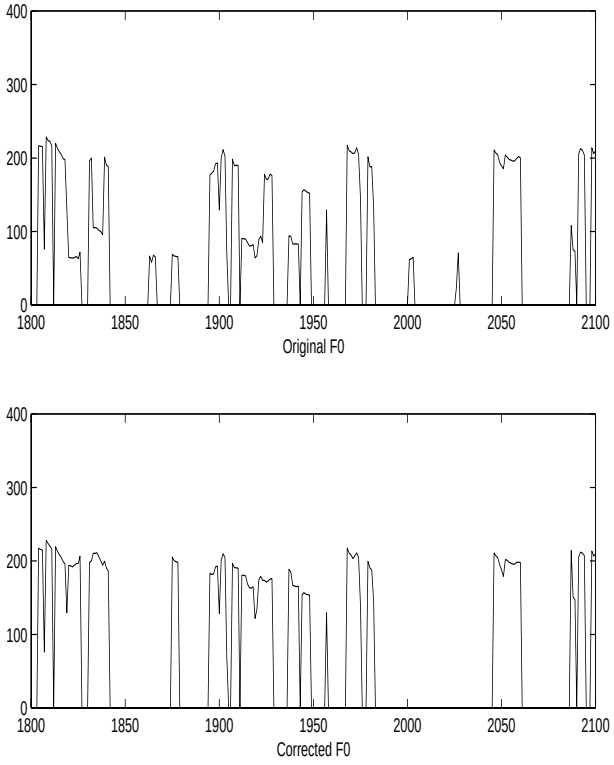


**Figure 1. Plots of F0 values before and after preprocessing.**

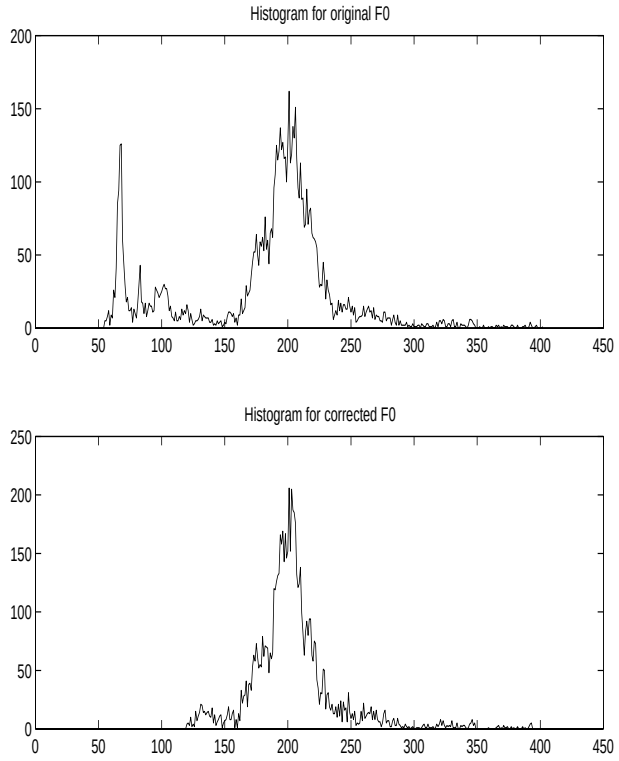


**Figure 2. Histogram of F0 values before and after processing for new estimate.**

represent the data [6].

### 3.1. Segmentation of F0 Trace for Analysis

The stretch of speech used as the speech analysis unit (SAU) over which the trend function is computed is defined by the user. In the processing described here, we use SAUs based on a pause analysis of F0 traces. The SAUs are defined as a stretch of speech between boundaries defined by the F0 pausal duration exceeding a threshold. SAU boundaries are not intonational phrase boundaries as the latter do not necessarily correspond to a pause. The SAUs turn out to be well-matched to linguistic units hand-marked by experts for investigating the intonation-gesture nexus.

Parameterized trend functions are fitted to the F0 data over the SAUs. The procedure allows a further segmentation of SAUs to fit trend functions over two sections within the SAU.

### 3.2. Computation of Parameters

The preprocessing described above is essential for obtaining parametric representations of F0 in SAUs. For each SAU, we consider a line segment or bar to represent the F0. This bar-fitting captures the average F0 value along with the low-resolution trend to determine whether the pitch increases, decreases, or remains constant. This information is useful in analysis, and also allows local fine variations to be evaluated after factoring out the low-resolution trends.

Since the data is noisy, the F0 trace contains substantial high frequency components. First the F0 trace is median-smoothed using a small window of size 3. We then find the line segment with slope $m$ and mid-interval value equal to $f_c$ that best fits the data according to a cost function. However the presence of unvoiced segments within a SAU have to taken into account when fitting line segments. The samples where F0 values are zero should be disregarded in computing the cost function.

Let $S$ denote the set of frame-indexes $\{x\}$ within a given SAU. Also let $f(x)$ denote the F0 value at index $x$. Define the mean index $x_c$ as the average of the values of $\{x\}$ in $S$. Let $S_v$ denote the voiced indexes where F0 is non-zero. Then we define a cost function $C(m, f_c)$ as

$$C(m, f_c) = \sum_{x \in S_v} g[f(x) - (f_c + m(x - x_c)]$$

The function g[.] is a non-negative function defined by

$$g[y] = \begin{cases} y^p & |y| \leq \Delta f, \\ (\Delta f)^p & |y| > \Delta f. \end{cases}$$

where we chose $p = 2$ in our computations.

The values of the slope $m$ and mid-interval value $f_c$ are chosen so that the cost function $C(m, f_c)$ is minimum. This procedure can be viewed as using a parallelogram in the plane of frame-index and F0 values, to fit a line-segment to the F0 data within a SAU. F0 values are modified to lie within the parallelogram. However when F0 values are zero, i.e. $f = 0$, then these are not included in computing the cost function, which uses mean-squared error criterion with saturation.

In some SAUs, a single bar cannot represent the F0 variation adequately. In such a case it is appropriate to partition the SAU into two segments and fit double bars. We use the same procedure as above, except that we change not only the height and slope, but also the break point between the two SAU segments. Now, the two line segments represent the F0 variation in one SAU.

The procedure for computing the parameters of the line segments for the case of both single and double segments was applied to a preprocessed F0 trace. The results are shown in Figure 3.

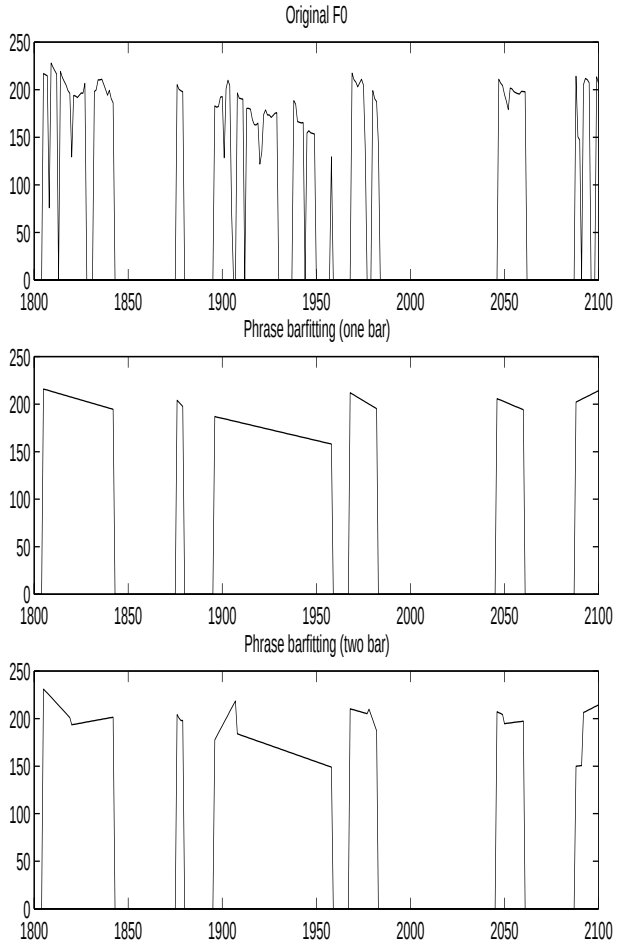


**Figure 3. The re-estimated F0, fitting a single bar, and a double bar, within segments**

## 4. DISCUSSION

In this paper we describe methods of processing the fundamental frequency (F0) trace in order to (i) obtain an improved estimate, (ii) segment the processed F0 data, and (iii) find a parametric representation suitable for the multimodal analysis. The choice of parametric representations can be varied within this framework. The goal is to use flexible parametric representations whose parameters can be examined for correlation with suitably extracted gesture features. An example of the multimodal primitives, i.e. hand position and F0, along with expert analysis is shown in Figure 4.

To study correlations, several alternate representations need to be examined. We note that the F0 trace consists of highly time-correlated data. The fact that discrete cosine transform (DCT) and discrete sine transform (DST) have excellent energy compaction for highly correlated data [7] makes them attractive for coding the speaker's prosody parameters. Based on our investigation [6], DCT was found to be suitable for coding pitch periods in isolated voiced segments. We are examining the use of DCT in representing F0 over a speech analysis unit as defined in section 3.

**Acknowledgment**

This work was supported the National Science Foundation under the grant IRI-9618887.

## REFERENCES

[1] G. Bailly and C. Benoit (Eds.), *Talking Machines, Theories,Models and Designs*, Elsevier, 1992.

[2] G. Brown, K. Currie, and J. Kenworthy, *Questions of Intonation*, University Park Press, Baltimore 1980.

[3] B. Grosz, J. Hirschberg, "Some Intonational Characteristics of Discourse Structure", *Proceedings of International Conference on Spoken Language Processing*, pp. 429-432, 1992.

[4] J. Hirschberg, B. Grosz, "Intonational Features of Local and Global Discourse Structure," 1997

[5] W.B. Kleijn and K.K. Paliwal (Eds.), *Speech Coding and Synthesis*, Elsevier, Amsterdam, 1995.

[6] W. Kurek, R. Ansari, "Speech Prosody Representation for Synthesis and Compression," *Proceedings of Second International Conference on Multimedia Information Systems*, Chicago, IL, pp. 220-225, April 1997.

[7] Anil K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989.

[8] D. R. Ladd, *Intonational Phonology*, Cambridge University Press, Cambridge, 1996.

[9] I. Lehiste, "Perception of Sentence and Paragraph Boundaries", *Frontiers of Speech Research*, London: Academic Press, pp. 191-201, 1979.

[10] M. Lieberman, Computer Speech Synthesis: Its Status and Prospects, *Voice communication between humans and machines*, National Academy of Sciences, 1994.

[11] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press, Chicago, 1992.

[12] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, NJ, 1978.

[13] D. B. Roe and J. G. Wilpon (Eds), *Voice communication between humans and machines*, National Academy of Sciences, 1994.

[14] R. W. Schafer, "Scientific Bases of Human-Machine Communication by Voice", *Voice communication between humans and machines*, National Academy of Sciences, 1994.

[15] A. Syrdal, R. Bennett, and S. Greenspan, *Applied Speech Technology*, CRC Press, Boca Raton, FL, 1995.

[16] J. Van Santen, R. Sproat, J. Olive, and J. Hirshberg, editors. *Progress in Speech Synthesis*, Springer Verlag, New York, 1995.

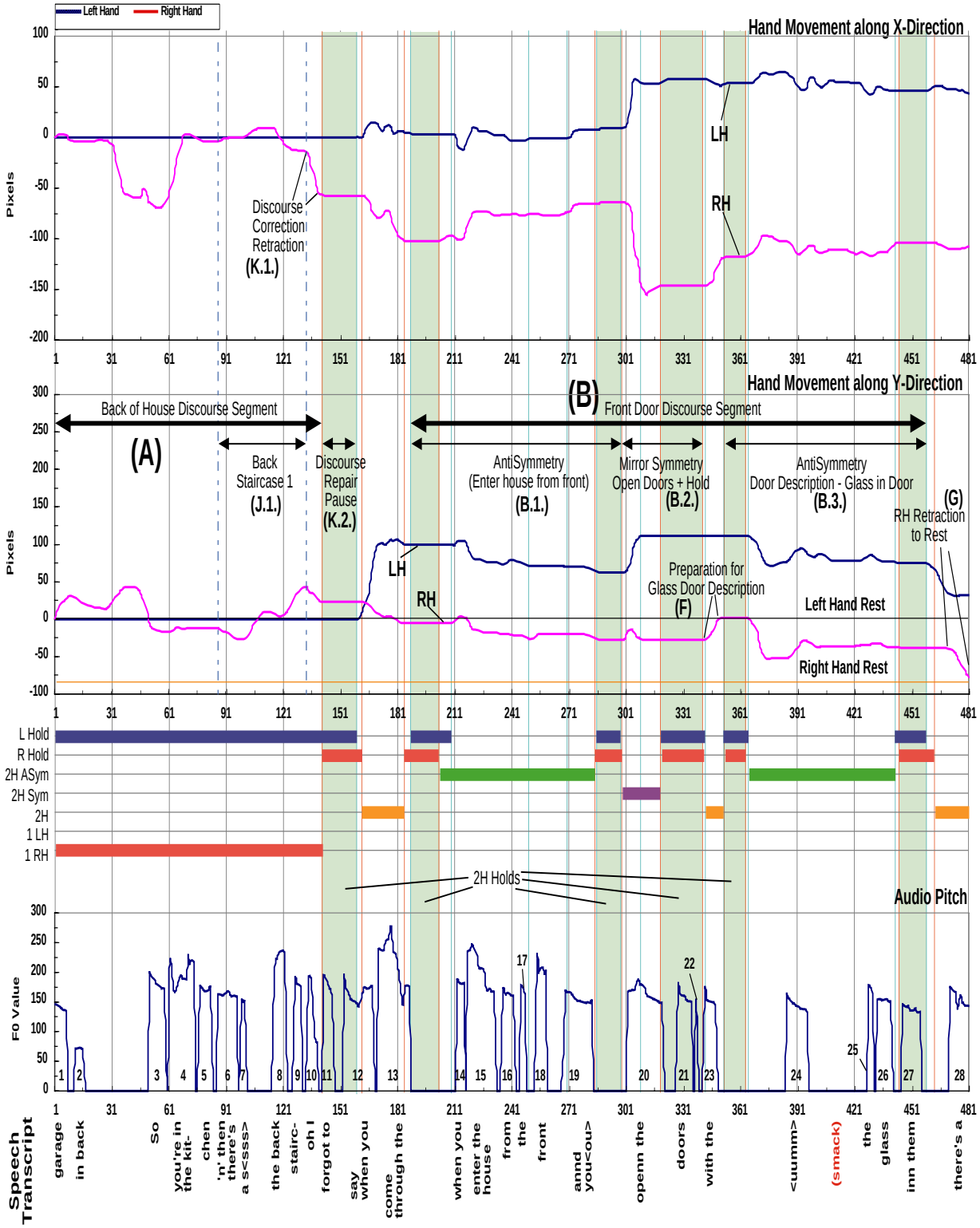


Figure 4. Gesture and speech: Plots of hand position, analysis, and F0 for frames 481-961.