Published in final edited form as:
*Speech Commun*. 2009 November 1; 51(11): 1082–1097. doi:10.1016/j.specom.2009.04.007.

# Automated Assessment of Prosody Production[1]

**Jan P. H. van Santen**, **Emily Tucker Prud'hommeaux**, and **Lois M. Black**
Center for Spoken Language Understanding, Division of Biomedical Computer Science, Oregon Health & Science University

## Abstract

Assessment of prosody is important for diagnosis and remediation of speech and language disorders, for diagnosis of neurological conditions, and for foreign language instruction. Current assessment is largely auditory-perceptual, which has obvious drawbacks; however, automation of assessment faces numerous obstacles. We propose methods for automatically assessing production of lexical stress, focus, phrasing, pragmatic style, and vocal affect. Speech was analyzed from children in six tasks designed to elicit specific prosodic contrasts. The methods involve dynamic and global features, using spectral, fundamental frequency, and temporal information. The automatically computed scores were validated against mean scores from judges who, in all but one task, listened to "prosodic minimal pairs" of recordings, each pair containing two utterances from the same child with approximately the same phonemic material but differing on a specific prosodic dimension, such as stress. The judges identified the prosodic categories of the two utterances and rated the strength of their contrast. For almost all tasks, we found that the automated scores correlated with the mean scores approximately as well as the judges' individual scores. Real-time scores assigned during examination – as is fairly typical in speech assessment – correlated substantially less than the automated scores with the mean scores.

## 1. Introduction

Assessment of prosody is important for diagnosis and remediation of speech and language disorders, for diagnosis of certain neurological conditions, as well as for foreign language instruction. This importance stems from the role prosody plays in speech intelligibility and comprehensibility (e.g., Wingfield, 1984; Silverman et al., 1993) and in social acceptance (e.g., McCann & Peppé, 2003; Peppé et al., 2006, 2007), and from prosodic deficits in certain neurological conditions (e.g., stroke; House, Rowe, & Standen, 1987 or Parkinson's Disease; Darley, Aronson, & Brown, 1969a, b; Le Dorze et al., 1998).

Current assessment of speech, including that of prosody, is largely auditory-perceptual. As noted by Kent (1996; also see Kreiman & Gerratt, 1997), the reliability and validity of auditory-perceptual methods is often lower than desirable as the result of multiple factors, such as the difficulty of judging one aspect of speech without interference from other aspects (e.g., nasality

[1]Preliminary results of this work were presented as posters at IMFAR 2007 and IMFAR 2008

Corresponding Author: Jan van Santen, Center for Spoken Language Understanding (CSLU), Division of Biomedical Computer Science (BMCS), School of Medicine, Oregon Health & Science University, 20000 NW Walker Road / Beaverton, OR 97006, Office: +1-503-748-1138 / Fax: +1-503-748-1306, vansanten@cslu.ogi.edu.

Disclosure: This manuscript has not been published in whole or substantial part by another publisher and is not currently under review by another journal.

judgments in the presence of varying degrees of hoarseness); the intrinsic multidimensional nature of certain judgment categories that require judges to weigh these dimensions (e.g., naturalness); the paucity of reference standards; and the difficulty of setting up truly "blind" judgment situations. Many of these issues are not specific to perceptual judgment of speech; in fact, there is an extensive body of literature on biases and inconsistencies in perceptual judgment going back several decades (e.g., Tversky, 1969).

Presumably, these issues would not be faced by automated ("instrumental") speech assessment methods. Nevertheless, automated methods have largely been confined to analysis of voice features that are only marginally relevant for prosody (e.g., the Multi-Dimensional Voice Program™, or MDVP; Elemetrics, 1993). What obstacles are standing in the way of developing reliable automated prosody assessment methods?

An obstacle for any method, whether automated or auditory-perceptual, consists of the multiple "levels" of prosodic variability; for each level, one must distinguish between which deviations from some – generally ill-defined – norm are acceptable (e.g., due to speaking style) and which deviations are not (e.g., due to disease). One level of variability involves dialect. Work by Grabe and colleagues (Grabe, Post, Nolan, and Farrar, 2000; Grabe and Post, 2002), for example, has shown that prosodic differences between dialects of British English can be as large as prosodic differences between languages. Not surprisingly, dialect differences have been shown to create problems for auditory-perceptual assessment (e.g., assessment of speech naturalness; Mackey, Finn, & Ingham, 1997). Another level of variability involves differences between speakers that are not obviously due to dialect. These differences are sufficiently systematic to provide useful cues for speaker identification (e.g., Sönmez et al., 1998; Adami et al., 2003), and may involve several speaker characteristics, the most obvious of which are gender, age, and social class (e.g., Milroy & Milroy, 1978). At a third level, there is systematic within-speaker variability due to task demands (e.g., Hirschberg, 1995, 2000), social context (e.g., Ofuka et al., 1994, 2000), and to emotional state (e.g., Scherer, 2003).

In addition to these forms of variability that are due to systematic factors, there is also variability that is apparently random. For example, in data reported by van Santen and Hirschberg (1994), in a highly confined task in which the speaker had to utter sentences of the type "Now I know <target word>" in a prosodically consistent – and carefully monitored – manner, the initial boundary tone was found to have a range of 30 Hz while the final boundary tone had a range of only 3 Hz; the typical pitch range of these utterances was less than 100 Hz. This form of variability may be less of a challenge for auditory-perceptual methods because these methods may benefit from the human speech perception system's ability to ignore communicatively irrelevant features of speech, but it clearly presents a challenge for automated methods.

There are additional aspects of prosody that pose complications and that are unrelated to variability. One is the intrinsically relative nature of many prosodic cues. For example, durational cues for the lexical stress status of a syllable are not in the form of some absolute duration but of how long or short the duration is compared to what can be expected based on the speaking rate, the segmental makeup of the syllable, and the location of the syllable in the word and phrase (van Santen, 1992; van Santen & Shih, 2000). A second aspect of prosody that poses complications for automated methods is that prosodic contrasts typically involve multiple acoustic features. To continue with the same example, lexical stress is expressed by a combination of duration, pitch, energy, spectral balance (e.g., Klatt, 1976; van Santen, 1992; Sluijter & van Heuven, 1996; van Santen & Niu, 2002, Miao et al., 2006), and additional features due to effects at the glottal level that are not fully captured by these basic acoustic features (e.g., glottal closing and opening slope; Marasek, 1996). Thus, there could be speaker-dependent trade-offs in terms of the relative strengths of these features, necessitating a fundamentally multidimensional approach to automated prosody assessment.

Both the intrinsic relativity of individual prosodic features and the trade-offs between them pose challenges for automated prosody assessment methods. These challenges seem to be fundamentally different from those posed by, for example, vowel production assessment. In a given phonemic context, vowel formant frequencies must lie within fairly narrow ranges in order for the vowel to be perceived as intended. While prosodic categories cannot even remotely be characterized by "point templates" in some conventional acoustic space, point template approaches for phonemic categories used by typical speech recognition systems clearly work rather well, via vector-based acoustic models in conjunction with some initial normalization step (e.g., cepstral normalization, vocal tract length normalization) and making basic allowances for coarticulation (e.g., by using phonemic-context dependent acoustic models).

Despite these obstacles for automated methods, there are obvious drawbacks to relying on auditory-perceptual methods and important advantages to using automated methods. First, we already mentioned validity and reliability issues of auditory-perceptual methods. Second, given the poor access many individuals have to services from speech-language pathologists or foreign language teachers, reliance on computerized prosody remediation or instruction is likely to increase. To be truly useful, such computerized systems should have the capability to provide accurate feedback; this, in turn, requires accurate automated assessment. Third, despite the exquisite sensitivity of human hearing, it is plausible that diagnostically relevant acoustic markers exist whose detection exceeds human capabilities. Detection of some promising markers, such as statistical features of pause durations in the course of a 5-minute speech recording (e.g., Roark et al., 2007), might be cognitively too demanding. Others could have too low an SNR to be humanly detectable. The acoustic feature of jitter, for example, has potential for early detection of certain vocal fold anomalies (e.g., Zhang & Jiang, 2008; Murry & Doherty, 1980) but has fairly high perceptual thresholds, certainly with non-stationary pitch (e.g., Cardozo & Ritsma, 1968). In other words, exclusive reliance on auditory-perceptual procedures is not good for discovery of new diagnostic markers.

We thus conclude that automated measures of assessment of prosody production are much needed, but that constructing such measures faces specific challenges. In our approach, we use a combination of the following design principles that help us address these challenges. (i) *Highly constraining elicitation methods* (e.g., repeating a particular word with a specific stress pattern) to reduce unwanted prosodic variability due to, for example, contextual effects on speaking style. (ii) *A "prosodic minimal pairs" design* for all but one task, in which the list of items used to elicit speech consists of randomized pairs that are identical except for the prosodic contrast (e.g., the third item on the list is tauveeb and the seventh tauveeb, with underlining indicating word stress). This serves to reduce the impact of confounding speaker characteristics, such as pitch range or vocal tract length; each speaker is his or her own control. (iii) *Robust acoustic features* that can handle, for example, mispronunciations and pitch tracking errors. (iv) *Measures that consist of weighted combinations of multiple, maximally independent acoustic features*, thereby allowing speakers to differ in the relative degrees to which they use these features. (v) *Measures that include both global and dynamic features*. Prosodic contrasts such as word stress are marked by pitch dynamics, while contrasts such as vocal affect can perhaps be characterized by global statistics. (vi) *Parameter-poor* (and even *–free*) *techniques* in which the algorithms themselves either are based on established facts about prosody (e.g., the phrase-final lengthening phenomenon) or are developed in exploratory analyses of a separate data set whose characteristics are quite different from the main data in terms of speakers (e.g., adults and children ages 11-65 vs. children 4-7). In conjunction with (ii) and (iii), this serves to maximize the portability of the measures in order to minimize the influences of recording conditions, SNR, sample characteristics, and other factors that may be difficult to control across laboratories or clinics. Parameter-rich systems may lack such

portability, since the parameter estimates may depend on the idiosyncrasies of the acoustic recording conditions and the training samples.

The goal of this paper is to describe the construction and validation of a number of prosody measures based on these design principles. The speech data were collected as part of an ongoing study of the production and interpretation of prosody in autism, whose aim is to detail prosodic difficulties in autism spectrum disorder, developmental language disorder, and typical development, in the age range of 4-8 years. The current paper focuses on methodology. Elsewhere we have presented preliminary findings on between-group differences on the suite of measures (Tucker Prud'hommeaux et al, 2008; van Santen, Tucker Prud'hommeaux, et al., 2007, 2008).

## 2. Data Collection

### 2.1 Speech Elicitation Methods

The tasks used for elicitation include variants and modifications of tasks in the PEPS-C ("Profiling Elements of Prosodic Systems - Children"; Peppé & McCann, 2003) paradigm, as well as of two tasks developed by Paul et al. (2005). In our study of prosody in autism, children complete tasks designed to test both their interpretation and their production of prosody. The present paper considers, in detail, the results of only the tasks related to the production of prosody. Findings are reviewed in four categories: *Stress Related Tasks*, *Phrasing Task*, *Affect Task*, *and Pragmatic Style Task*.

*(i) Stress Related Tasks (Lexical Stress Task, Emphatic Stress Task, and Focus Task)*. In the *Lexical Stress Task* (based on Paul et al., 2005; also see Dollaghan & Campbell, 1998), the computer plays a recording of a two-syllable nonsense word[2] such as tauveeb, playfully accompanied by a picture of a thus-named "Martian" life-form. The child's task is to repeat after the recorded voice with the same stress pattern. In the *Emphatic Stress Task* (Plant & Öster, 1986; Shriberg et al., 2001; Shriberg et al., 2006) the child has to repeat an utterance in which one word is emphasized ("Bob may go home", "Bob may go home", etc.). (Note that in Plant & Öster's procedure, the subject does not repeat an utterance but reads appropriately annotated text aloud.) Finally, in the *Focus Task* (adapted from the PEPS-C), a recorded voice incorrectly describes a picture of a brightly colored animal with a soccer ball, using the wrong word either for the animal or for the color. The child must correct the computer by putting contrastive stress on the incorrect label. For example, if the voice describes a picture of a black cow as a blue cow, the child responds "No, the black cow has the ball". (ii) In the *Phrasing Task*, also adapted from the PEPS-C, the child has to indicate with an appropriate phrase break whether a picture represents three objects (e.g., fruit, salad, and milk) or two objects (e.g., fruit-salad and milk). (iii) In the *Affect Task*, the child has to say a fixed phrase ("It doesn't matter") using the emotion (happiness, sadness, anger, or fear) that corresponds to the affect expressed by a picture of a stylized face. These pictures were obtained by digitally modifying Ekman faces (Ekman and Friesen, 1976) into line drawings that only retain features relevant for facial affect. This task is loosely based on a receptive vocal affect task (Berk, Doehring, & Bryans, 1983). (iv) In the *Pragmatic Style Task* (based on Paul et al., 2005), the child views a photo of either a baby or an adult and must speak to that person using the appropriate prosody.

These tasks were administered as follows. Each task started with four training trials during which the examiner corrected and, if necessary, modeled the response. In addition, each task

[2]We use nonsense words rather than attested trochaic/iambic pairs (e.g., contest-N vs. contest-V) because such pairs are not identical phonetically and are not generally part of a typical child's vocabulary. We do recognize, as one reviewer pointed out, that inherent properties of the component phonemes may trigger the perception of stress in the absence of meaning. The minimal-pair presentation should reduce the likelihood of this phenomenon.

was immediately preceded by the corresponding receptive task, thereby providing some degree of additional, "implicit" training. Thus, significant efforts were made to ensure that the child understood the task requirements. During the administration of a task, the child and the examiner were seated at adjacent sides of a low, child-appropriate, rectangular table. The examiner interacted with a laptop computer to control the experiment; the screen of the laptop was not visible to the child. A software package (based on the CSLU Toolkit; Sutton et al., 1998) presented the auditory stimuli via high-quality loudspeakers and visual stimuli on a touch screen, processed the child's touch screen responses, and enabled to examiner to control the sequence of stimuli (specifically, to repeat up to two times items missed due to the child's attention drifting) and to indicate whether the child's response was correct or incorrect. The binary (i.e., correct vs. incorrect) scores obtained in this manner will be called *real-time scores*. The software also recorded the child's vocal responses and stored these in a data structure containing all events and speech recordings, appropriately synchronized and time-stamped. This data structure, including the recorded speech, can be re-accessed by the examiner after completion of a task to verify the scoring judgments made during the task. We call these scores *verified real-time scores*. The only real-time scores we will consider in this study are these verified real-time scores.

### 2.2 Speakers

Participants were 15 children who met criteria for Autism Spectrum Disorder (ASD); 13 children who were considered "typical" in terms of several criteria, discussed below (Typical Development, or TD group); and 15 children who met some but not all criteria for inclusion in the ASD group. There are two reasons for using this heterogeneous sample. First, for the purpose of developing and validating the automated measures, a wide range of performance levels are needed; restriction of the study to the TD group could have created serious restriction of range problems because most children in the TD group perform well on these tasks. Second, it is important to establish that the experimental procedures can be applied to, and generate meaningful scores for, children with neurodevelopmental disorders.

All participants were verbal and had a mean length of utterance of at least 4. The ASD group was "high functioning." (HFA), with full scale IQ of at least 70 (Gilberg & Ehlers, 1998; Siegel, Minshew & Goldstein, 1996), as well as "verbally fluent." Diagnoses made use of DSM-IV-TR criteria (DSM-IV-TR, 2000) in conjunction with results of the "revised algorithm." derived from the ADOS (Lord et al., 2000; Gotham, et al., 2007, 2008), administered and scored by trained, certified clinicians, and results of the Social Communication Questionnaire (SCQ; Berument et al., 1999).

Exclusion criteria for all groups were the presence of any known brain lesion or neurological condition, including cerebral palsy, tuberous sclerosis, intraventricular hemorrhage in prematurity, the presence of any "hard" neurological sign (e.g., ataxia); orofacial abnormalities (e.g., cleft palate); bilinguality; severe intelligibility impairment; gross sensory or motor impairments; and identified mental retardation. For the TD group, exclusion criteria include, in addition, a history of psychiatric disturbance (e.g., ADHD, Anxiety Disorder), or any family member with ASD or Developmental Language Disorder.

In addition to the child speakers, we also obtained recordings of a superset of the items in the above six tasks from 36 professional and amateur actors, both male and female, ranging in age from 11 to 65. We have used these recordings as a development set for exploratory purposes. Results for one of these individuals, for the affect task, were reported by Klabbers et al. (2007).

## 2.3 Listeners

Two sets of human listener-judged scores were collected: (1) verified real-time scores, produced by clinicians during examination and later verified; and (2) listener judgments, collected using a web-based randomized perceptual experiment, produced by six naïve, non-clinical individuals.

**2.3.1 Clinical Examination—**The four individuals generating the verified real-time scores had clinical or research experience with neurodevelopmental disorders and were extensively trained in administering and scoring the speech elicitation tasks (section 2.1). Their backgrounds were in speech language pathology, phonetics, and clinical child psychology. All were native speakers of American English. Since they tested different subsets of children, no mean examiner scores could be computed. Each of these individuals verified his or her own scores, off-line.

**2.3.2 Listening Experiment—**Six individuals participated in the listening experiment portion of the study. They reported normal or corrected-to-normal hearing, were native speakers of American English, and had no clinical or research experience with neurodevelopmental disorders. They were unfamiliar with the study goals and did not know the children. They were paid for their participation. They processed the same sets of stimuli, thus enabling us to compute mean listener scores.

## 2.4 Listening Experiment Tasks

All listener tasks were computer controlled and took place in quiet offices using headsets. In the Affect Task, a listener heard a recording of a child response, decided which of four alternative emotions (happy, sad, angry, or fearful) best described it, and indicated a certainty level on a 0-2 scale (0=possibly, 1=probably, 2=certainly). The four alternatives were indicated using the same set of stylized faces that was used in the task performed by the children, with descriptive verbal labels added. The locations of the four alternatives remained constant throughout the experiment. Listeners also had the option to indicate “I don't know” without selecting a degree of certainty. This option was included in response to listener requests for such an option during a pilot study.

In the remaining five tasks, a listener heard a recording of two child responses forming a *prosodic minimal pair* (e.g., tauveeb and tauveeb, from the same child), decided which member of the pair corresponded to which response alternative, and indicated one of two degrees of certainty (“probably”, “certainly”). Listeners also had the option to indicate “I don't know”. Scores generated by these procedures will be called *minimal pairs based scores*. While the certainty ratings add to task complexity, listeners expressed the need to be able express the sharp differences in confidence they experienced.

## 2.5 Task Scoring

### 2.5.1 Listening Experiment Scores

**Affect Task:** The listeners had substantial difficulty reliably distinguishing between angry and happy and between sad and fearful. Based on this, we scored the listener responses as follows. We scored a response (to an individual utterance) as positive when the listener chose happy or angry, with full confidence scored as 1, some confidence as 0.66, and a little confidence as 0.33; sad or fearful responses were scored likewise but with corresponding negative values. A response of “I don't know” was scored as 0. We note that this scoring scheme does not take into account what the utterances' intended affects are. Thus, correlations between judges based on these per-utterance scores do not reflect agreement about the *appropriateness* of the utterances. Recall that each speaker produced 4 versions of a single sentence: happy, sad, angry,

and fearful. To obtain per-speaker scores, we combined the scores assigned to that speaker's utterances by adding the mean listener scores for the utterances whose targets were angry or happy and subtracting from this sum the sum of the scores for the utterances whose targets were sad or fearful. This difference score, in contrast to the per-utterance score, does reflect whether a child makes an appropriate contrast between the affects.

**Remaining five tasks:** In these tasks, in which the listener judged minimal pairs of utterances, we assign a positive score to a listener response if the listener's choice indicates that a child made the correct contrast between the two utterances and a negative score otherwise. A score of zero indicates that neither utterance was produced with convincingly appropriate prosody. Thus, scores could have the values of 1, 0.5, 0, -0.5, and -1. Per-speaker scores were computed by averaging these scores.

**2.5.2 Verified Real-Time Scores—**During an examination, the clinician – who always knew the target prosody – decided whether each response was correct or incorrect, without a certainty rating (correct = 1, incorrect = 0). During verification, the clinician listened to each response and decided whether to accept or reject the decision made during the examination. To make these scores comparable to the minimal pairs based scores, we collected for each minimal pair the two corresponding verified real-time scores, and mapped these onto 1.0, 0, and -1.0 depending on whether both were positive, one was positive, or neither was positive.

## 3. Listening Task Results

For each task, the listener data can be represented as an N × 6 table, where the N rows correspond to the utterances collected across all speakers, the 6 columns to the listeners, and the cells to the listener ratings. This *per-utterance* table can be condensed into a *per-speaker* table, by combining for each speaker the ratings of the *k* utterances as discussed in section 2.5. We extend these tables by adding additional columns for mean listener scores (see below) and verified real-time scores (see Section 2.1).

Two types of mean listener scores were computed. The first type of mean listener score was obtained simply by averaging the six listener columns in the data tables. The second score was obtained by applying Principal Components Analysis to the covariance matrix of the six listener columns, and multiplying these columns with the Eigen vector associated with the first principal component. This score is thus a weighted mean of the listener scores, with lower weights for listeners whose scores are less correlated with other listeners (van Santen, 1993). Since we found only minimal differences between the results obtained by these two methods, we only report results obtained with the simple average. Consistent with these minimal differences is the fact that, across the six tasks, the first component explained between 63 and 78% of the variance for the per-utterance data and between 76 and 90% of the variance of the per-subject data. There were no systematic differences between listeners in terms of their loadings on the first principal component. In combination with the similarity of the results obtained between weighted mean listener scores and the simple mean listener scores, we conclude that there were no systematically distinct (“bad”) subgroups of listeners or individual outliers and that we can use the simple mean listener scores.

Figures 3-8, left panels, show the (product-moment) correlations between the scores of the per-utterance listener scores, mean listener scores, and verified real-time scores. The figures show the ranges of the between-listener correlations and of the correlations between the listeners and the mean listener scores; for the latter, we always excluded the individual listener's data when the mean listener score was computed to avoid the obvious inflation that would otherwise result. Figures 3-8, right panels, present the same analyses for per-speaker data. One could

propose that from a practical perspective, it would be these per-subject scores that matter and not the per-utterance scores.

These results show, first, that all scores have significantly positive correlations with each other (p=0.01, one-tailed). Second, the listener scores correlate more with each other (on average, 0.64 for the per-utterance data and 0.73 for the per-speaker data) and with the mean listener scores (0.75 and 0.83, respectively) than the verified real-time scores (0.48 and 0.56, respectively).

Several factors could be responsible for the latter, including: frame-of-reference effects due to different subgroups of children being scored by different examiners; the examiners being biased by awareness of the broad capabilities of the child they were scoring; and judgments of individual responses being intrinsically more difficult than judgments of minimal pairs.

Except for the affect task, all correlations were larger for the per-speaker scores than for the per-utterance scores. This was most likely due to the reduction in variance by computing per-speaker averages. In the case of the affect scores, the per-utterance and the per-speaker scores are not strictly comparable because the former do not take into account the appropriateness of the utterance whereas the latter do.

A finding not depicted in the figures is that mean listener scores on the tasks other than the affect task were overwhelmingly positive (i.e., indicating that the child made the correct distinction), whether based on per-utterance data or on per-speaker data. Averaged over these tasks, the percentages of positive responses were 88% and 95%, respectively. This makes an important point: even in cases where the listeners were quite uncertain, as a group they were still able to accurately determine the child's intention. This supports the sensitivity of the prosodic minimal pairs-based methods. The verified real-time scores were less often positive (78% and 89% for the per-utterance and per-speaker data, respectively; using a chi-square test, these percentages are significantly smaller at $p<0.01$ than the corresponding percentages for the minimal pairs based methods). A methodological consequence of the strong positive bias is that the percentage agreement between judges (i.e., agreement measured in terms of the sign of the scores, ignoring the magnitude) is not meaningful, and that the standard measure for interjudge agreement (Cohen's Kappa; Cohen, 1960) is also not meaningful because this measure becomes unstable as the number of cases in the negative/negative cell approaches zero.

In summary, the data support the claim that the mean listener scores can serve as a gold standard for validating the automated scores because of the good agreement of the individual listeners' scores with each other and with the mean listener scores, and because of the sensitivity of these scores to the children's intended prosodic contrasts. The data show that the verified real-time scores are less reliable, and hence should not be used to evaluate the automated scores.

## 4. Description of Automated Measures

### 4.1 Pre-processing

Pre-processing consisted of the following steps: locating and extracting the child responses from the recordings; determining certain syllable, word, or phonetic segment boundaries (depending on the task and measure used; discussed in detail below); and extracting acoustic features, including (i) fundamental frequency, or $F_0$, using the Snack Sound Toolkit (2006), and (ii) amplitude (in dB) in four formant range based passbands ($B_1(t)$: 60-1200 Hz, $B_2(t)$: 1200-3100 Hz, $B_3(t)$: 3100-4000 Hz, $B_4(t)$: 4000-8000 Hz). These passbands are based on Lee et al. (1999, Table III), using formant ranges for 5-year olds. The amplitude trajectories $B_1(t)$, …, $B_4(t)$ are transformed into two new measures: $B(t)$, the amplitude averaged over the

four passbands, defined as $0.25*[B_1(t)+ \ldots + B_4(t)]$; and a *spectral balance vector* consisting of $b_j(t)$, the mean-corrected energies, defined as $b_j(t)=B_j(t)-B(t)$, for $j=1,\ldots,4$.

Spectral balance, thus defined, has a number of useful properties. (i) It is gain invariant, i.e., multiplication of the input signal by a constant has no effect. To a first order of approximation, this makes these features robust with respect to factors such as the gain setting of the hardware and the proximity of the speaker to the microphone. The same is obviously not true for $B(t)$. However, the prosodic minimal pairs approach can be expected to reduce the effects of these factors since they do not vary substantially within a session and hence approximately have the same values in the two utterances making up a prosodic minimal pair. (ii) Spectral balance should not be confused with spectral tilt defined as, e.g., the slope of a line fitted to the log power spectrum. As we shall see in the analysis of affect in Section 4.8, certain affects are associated with high values of $b_1$ and $b_4$ and low values of $b_2$ and $b_3$, while other affects have the opposite pattern. This difference cannot be captured by spectral tilt. (iii) The advantage of using formant frequency-based passbands is that it reduces effects on these measures of articulatory variability. If nominally the same vowel is pronounced differently in the two utterances making up a prosodic minimal pair, resulting in different formant frequencies, this will have limited effects on the spectral balance vector because of the way we selected the passbands.

### 4.2 Stress Related Tasks (Lexical Stress Task, Emphatic Stress Task, Focus Task): "Dynamic Difference" Based Measures

Broadly speaking, in stress related tasks the expected contrast involves a change in the *alignment* of an acoustic trajectory relative to syllables or words. For example, the $F_0$ peaks in tauveeb and tauveeb might be expected to occur in the first and second syllable, respectively. Alignment, however, is not nearly as simple. A method that considers which syllable contains a pitch peak will not work because $F_0$ peaks can occur in the stressed or post-stressed syllable depending on the segmental make-up of the syllables and the number of syllables in the (left-headed) foot with which the pitch accent is associated (van Santen & Möbius, 2000; van Santen, Klabbers, & Mishra, 2006). An additional problem is the difficulty of determining in which syllable the $F_0$ peak is located, due to the existence of a flat region in the $F_0$ curve, to the presence of small amounts of jitter, and to the "true" $F_0$ peak being located inside an obstruent where pitch cannot be measured. Our method, instead, uses a "soft" alignment approach in which we (i) compute a *difference curve* between the two curves extracted from the utterances making up a prosodic minimal pair and (ii) use a one-dimensional *ordinal pattern recognition* method based on *isotonic regression* (Barlow et al., 1972) to determine whether the difference curve is closer to the *ordinal pattern* consistent with a correct response pair than to the ordinal pattern associated with an incorrect response pair. The straightforward intuition behind this concept is that even under considerable variability in alignment, this difference curve has an up-down-up shape (see Figure 1).

We first explain the ordinal pattern detection method in general form. Consider a sequence of values, $x=<x_1, \ldots, x_n>$, and associated weights, $w=<w_1, \ldots, w_n>$. The $x$'s, for example, may denote fundamental frequency values sampled at 5 ms frame intervals and the $w$'s per-frame products of amplitude and voicing probability. An *ordinal pattern* is defined in terms of sequences of substrings $<1,\ldots,i_1>, <i_1+1,\ldots,i_2>, \ldots, <i_{m-1}+1,n>$ of the string $<1,\ldots,n>$ together with *directions* denoted *u* (for up) or *d* (for down) associated with these substrings. We define *the fit to an ordinal pattern udu*… as follows.

$$Fit(x, w, udu\ldots)=\min_{\substack{1 \le i_1 \le \cdots \le i_{m-1} \le m \\ y_1 \le \cdots \le y_{i_1} \\ y_{i_1}+1 \ge \cdots \ge y_{i_2} \\ \cdots \\ y_{i_{m-1}}+1 \le \cdots \le y_m}} \sum_{k=1,\ldots,m-1} \sum_{j=i_{k-1}+1,\ldots,i_k} w_j(x_j - y_j)^2 \quad (1)$$

Similarly for a pattern *dud*…, with appropriate changes in the inequalities. For the special case of *Fit(x,w,u)*, this equation reduces to the standard isotonic regression (Barlow et al., 1972), given by:

$$Fit(x, w, u)=\min_{y_1 \le \cdots \le y m_1} \sum_{j=1,\cdots,m} w_j(x_j - y_j)^2 \quad (2)$$

Thus, the measure of fit in Eq. (1) alternatingly applies isotonic regression and antitonic regression (in which the ≤ signs are replaced by ≥ signs) for any selection of turning points, $i_1,\ldots,i_{m-1}$, and jointly optimizes the isotonic and antitonic fits over all selections of these turning points.

The *Dynamic Difference* method comprises the following steps. Below, $s_L(t)$ and $s_R(t)$ are the trajectories that, if the correct prosodic contrast is made, are associated with the *left aligned item* (e.g., tauveeb) and *right aligned item* (e.g., tauveeb), respectively:

1. Time-warp $s_L(t)$ such that the phonetic segment boundaries coincide with those associated with $s_R(t)$.
2. Compute the difference curve, $s_L(d[t]) - s_R(d[t])$, where *d[]* is the time-warp.
3. Using Eq. (1), compute the fits of $f_{dud\ldots} = Fit[s_L(d[t])- s_R(d[t]), w, udu\ldots]$ and the fit to the *converse* of *udu*…, $f_{dud\ldots}=Fit[s_L(d[t])-s_R(d[t]), w, dud\ldots]$. For weights, $w_j$, we use the product of per frame amplitude and voicing probability, as computed by ESPS's *get_fo* program (Entropic, 1996), implemented in the Snack package (The Snack Sound Toolkit, 2006), in the case of pitch analysis, and voicing probability-only in the case of the energy based acoustic features, $b_j(t)$ and $B(t)$.
4. Compute $(f_{dud\ldots}-f_{udu\ldots}) / (f_{dud\ldots} +f_{udu\ldots})$. This measure has a value of +1 (-1) when the fit to *udu*… (*dud*…) is perfect, and has values in between -1 and +1 depending on which fit is better. We call this measure the *dynamic difference between $s_L$ and $s_R$*.

We computed two dynamic difference measures in the case of the Lexical Stress Task *(dud/udu* and *d/u* based), but only the *d/u* based measure for the Emphatic Stress Task and the Focus Task (in which cases we analyzed the inter-vowel-center intervals); the key reason is that the pitch movement is confined to the critical word in the first task but also includes utterance parts subsequent to the critical word pair in the latter tasks; these parts are generally not well-controlled by the child and hence introduce severe phonemic variability (e.g., the insertion of entire words) that undermines the comparability of the utterances making up a prosodic minimal pair. For consistency, we only report results from the *d/u* based analyses. (For the Lexical Stress Task, results were essentially the same for the *dud/udu* and *d/u* based analyses.)

### 4.3 Stress Related Tasks (Lexical Stress Task, Emphatic Stress Task, Focus Task): Duration based measure

It is well-known that word stress and sentence stress are associated with an increase in duration of the stressed syllable or word (Klatt, 1976). We propose to use a simple measure, given by $(L_1R_2-L_2R_1,)/(L_1R_2+L_2R_1)$, where $L_i$ and $R_i$ denote the duration of the $i$-th syllable or word in the left (L) and right (R) aligned items, respectively.

### 4.4 Stress Related Tasks (Lexical Stress Task, Emphatic Stress Task, Focus Task): Lexical Stress Ratio based measure

The Lexical Stress Ratio (*LSR*; Shriberg et al., 2003) was designed in the context of assessing the ability to convey lexical stress in trochaic words in children with childhood apraxia of speech. This measure is computed as follows (see Shriberg et al., 2003, for details). For a two-syllable word, *w*, with stress on the first syllable, the vowel regions are determined in each syllable, and pitch (in Hz) and amplitude are computed. Next, for each of the two syllables ($i$=1,2) the quantities $D_{wi}A_{wi}$ (*amplitude area*), $D_{wi}F_{0wi}$ (*pitch area*), and $D_{wi}$ are computed, where, for the $i$-th syllable, $D_{wi}$ is the duration of the vowel (in ms), $A_{wi}$ the average vowel amplitude (in dB), and $F_{0wi}$ the average fundamental frequency (in Hz) in the vowel region. In the next step, the ratios $D_{w1}A_{w1}/D_{w2}A_{w2}$, $D_{w1}F_{0w1}/D_{w2}F_{0w2}$, and $D_{w1}/D_{w2}$ are formed. Finally, these three ratios are combined via

$$\begin{aligned} LSR_w = \alpha\ & D_{w1}A_{w1} \\ & /D_{w2}A_{w2} + \beta D_{w1}F_{0w1} \\ & /D_{w2}F_{0w2} + \gamma D_{w1}/D_{w2}, \end{aligned} \tag{3}$$

where α, $\beta$, and $\gamma$ were determined by Shriberg et al. (2003) by computing the weights of the factor explaining the largest amount of variance, as produced by factor analysis (applied to a data matrix with speakers as rows, the three ratios as columns, and the per-speaker averages of these rows in the cells). The values of the weights were 0.507, 0.490, and 0.303, respectively. To apply this measure to the prosodic minimal pairs setup, we compute for each word pair ($L$, $R$), where L (R) denotes the left (right) aligned item, the measure $(LSR_L-LSR_R)/(LSR_L+LSR_R)$.

### 4.5 Stress Related Tasks: Non-minimal pairs analysis

To illustrate the value of a minimal pairs based analysis for reducing the impact of confounding speaker variables, we correlated the verified real-time scores with pitch, amplitude, and duration measures applied to individual utterances. These analyses were performed for the lexical stress task.

For pitch and amplitude, we applied the *d/u* based dynamic difference measure, assuming that when the pitch peak is on the first (second) syllable or word the overall trend in the two-syllable sequence is downward (upward). We also measured the ratio of the durations of the first and second syllable, assuming again that stress causes lengthening. A fundamental problem with these measures is, of course, that they are affected not only by speaker variables (e.g., some children with language disorders are very insecure about the correctness of their answer and express this with a rising intonation, a trend that is partialed out in the minimal pairs analysis but not in the current analysis) but also by the segmental structure of the syllables and words. For example, the second syllable in *shinaig* contains an intrinsically long diphthong; in addition, because this non-word is spoken in isolation, one may expect utterance final lengthening of the second syllable. These and other factors may conspire to make the proposed,

individual-utterance based measures unreliable as a way of assessing the correctness and strengths of prosodic contrasts.

We computed two types of correlations. First, we correlated the measures with the prosodic targets, scoring utterances with target stress on the first (second) syllable as -1 (+1). The correlations with the pitch, amplitude, and duration-based measures were 0.642, 0.527, and 0.235, respectively; the multiple regression correlation was 0.688. (Since the dependent variable was binary, this analysis is equivalent to linear discriminant analysis.) These correlations were similar to those between the mean listener scores and the corresponding minimal pairs based measures, although the multiple regression correlation was substantially higher in the latter case (0.79). These results show that, as noted before in Section 3 based on the real-time scores and on the listener scores, the children generally made appropriate prosodic contrasts. The results also show that the proposed features may be more useful for prosody recognition identification purposes than anticipated.

Second, we correlated the measures with the verified real-time scores. For this purpose, we first separated the utterances into two groups in accordance with the stress location of the target prosody. We then reversed the signs of the measures for the group with target stress on the second syllable. We now correlated these measures with the correct (1) vs. incorrect (o) verified real-time scores. For the first syllable stress group the correlations with the three measures were 0.125, 0.145, and 0.235, while for the second syllable stress group they were 0.338, 0.342, and 0.210. When the two groups we combined, the correlations were 0.291, 0.212, and 0.195. These results suggest that the correct/incorrect decisions correlated rather poorly with the proposed measures.

In summary, the proposed measures were moderately accurate in identifying the target prosody of an utterance. We hypothesize that this is because the children generally conveyed the target prosody clearly and correctly. However, the measures were not able to predict when an examiner would judge an utterance as being correct or incorrect.

### 4.6 Phrase Boundary Task: Duration based measure

A measure mathematically similar to the duration based measure in Section 4.3 was used, comparing the duration of the first word (e.g., “chocolate”) and the duration of the remainder (“cookies and jam”), with the boundary drawn at the start of the second word (“cookies”). In the case of “chocolate, cookies, and jam”, $P_1$ (the duration of the first word in the phrase boundary condition) is relatively long because it includes the full word “chocolate”, with the final syllable being lengthened by the phrase boundary (e.g., Klatt, 1976), and any pause; in the case of “chocolate-cookies and jam”, $N_1$ (the duration of the first word in the no-phrase boundary condition) is relatively short because the word “chocolate” is often contracted (e.g., to “chocl”) and there is neither phrase-final lengthening nor a pause. Since there is no reason why there would be substantial effects of the phrase boundary on the duration of the remainder (“cookies and jam”), $P_2$ and $N_2$ can be expected to be approximately equal and hence $P_1/P_{2\ to}$ be larger than $N_1/N_2$. A measure confined to the [-1,1] interval is given by $(P_1N_2\text{-}P_2N_1)/(P_1N_2+P_2N_1)$.

We spent substantial time exploring effects on pitch, but no consistent patterns were found. We surmise that this may be related to the fact that, while each of the prosodic minimal pairs in this task indeed involves the presence or absence of a phrase boundary, the linguistic nature of the boundary varies significantly across the items. In the two-item list context, each of the initial items (“chocolate cookies,” “chocolate ice cream,” and “fruit salad”) has a different sequence of grammatical components (ADJ-N, ADJ-N-N, N-N), each of which is associated with one or more default stress patterns. We also observed that some children emphasized “chocolate” and “fruit” in the two-item list context, perhaps to distinguish the pictured cookies,

ice cream, and salad from the sugar cookies, vanilla ice cream, and vegetable salad that they remembered from previous pictures depicting three items. We speculate that these somewhat subtle distinctions between items may further add to individual differences and hence to the lack of consistent pitch patterns for this task, because some children in this age range (and populations studied) may be able to understand and prosodically express these subtler differences and others do not.

### 4.7 Pragmatic Style Task: Global measures

Based on analyses of the actor data, we compute the following features. First, the average over the utterance of the fourth amplitude band, $B_4(t)$. We surmise that this band captures the "breathiness" of baby-directed speech which we informally observed in the actor data. Second, we compute a robust maximum value of $F_0$ by ordering all frames in the utterance in terms of the weight function defined by the product of amplitude and voicing probability, and finding the frame that has the largest $F_0$ value among the top 10% of the weight-ordered frames. Duration was not a reliable predictor of Style in the actor data and was not included in the evaluation.

### 4.8 Affect Task: Global measures

Based on the analysis of the actor data, we decided to include the (b2+b3)-(b1+b4) contrast (see Figure 2), which captures a striking pattern whereby the angry and happy affects have an inverted-U-shaped pattern in terms of the four bands while the sad and fearful affects have the inverse pattern. We also included the same robust $F_0$ maximum as well as the mean amplitude (via B1+B2+B3+B4).

## 5. Evaluation of Automated Measures

The key evaluation criterion is the degree to which the objective measures, separately and in combination, can predict the mean listener scores, as measured by the product-moment correlation. There are many ways in which, for a given task, the multiple acoustic measures can be combined, including neural nets, support vector machines, and other linear or non-linear methods. We have decided to use simple linear regression because of its robustness and its small number of parameters, which may benefit portability. To avoid the positive bias in estimating the multiple regression's R-squared statistic, we use a training/test procedure in which we estimate the regression weights on a subset of the data (training set) and evaluate the multiple correlations on the remaining data (test set). At the same time, in order to reduce restriction-of-range artifacts that could lead to underestimates of the true correlations between observed and predicted mean listener scores, we select training and tests sets with a sub-sampling procedure whose aim is to minimize differences in variance between training set, test set, and the overall set. Toward this end, we construct the test and training subsets by rank-ordering the data on the mean listener scores, and creating $n$ pairs of tests sets and training sets, $(Te_i,Tr_i)$, where the $i$-th test set, $Te_i$, contains items $i$, $n+i$, $2n+i$, …, and the corresponding training set, $Tr_i$, the remaining items. Typically, we use n=10. This results in the test sets and training sets having approximately the same variances as the entire data sample. We report correlations between automated scores and mean listener scores by computing the median of these correlations over multiple selections of test sets.

Results are shown in Figures 3-8, with the left panels for the per-utterance scores and the right panels for the per-speaker scores. We note the following:

**i.** As measured by the correlations with the mean listener scores, the proposed measures based on multiple regression (called "combined" in the figures) outperform the verified real-time scores, and, except for the per-utterance affect data and the per-speaker EST data, are within the range of correlations between individual listeners

and the mean listener scores. In several cases, even single measures perform at this level (e.g., the dynamic difference based $F_0$ measure in the Lexical Stress Task).

Thus, we have demonstrated the fundamental feasibility of developing automated measures whose reliability is comparable to and perhaps superior to manual scores.

**ii.** The LSR- based measure (Figure 4) performs less well than the dynamic difference based measure. We conjecture that this is related to two features of the LSR. First, weights produced by factor analysis do not optimize the ability to discriminate between words with strong vs. weak stress on the first syllable relative to the second syllable. Instead, factor analysis maximizes the variance of the children on the LSR measure. It is theoretically possible that the children vary on a dimension orthogonal to stress, such as the degree to which some children use duration while others use pitch to convey stress; under certain conditions, this pattern could result in the first factor being unrelated to stress. Second, when we inspect the predictive values of the area-based measures that are combined into the LSR, we see that these measures are outperformed by the dynamic difference-based measures. This may be due to the fact that using area-based – and hence average-value-times-duration-based – measures puts too much emphasis on duration (the three measures correlate in excess of 0.90), while we see in all stress-related tasks that duration is generally a weaker predictor than $F_0$.

**iii.** The Focus and EST tasks (Figures 3 and 5) showed similar patterns of results, in which $F_0$ is generally the strongest cue and amplitude the weakest. This similarity could suggest that the output demands of these two tasks are similar.

**iv.** It is of interest to revisit the issue of trade-offs between acoustic cues in the expression of stress. We note that correlations between the three measures used to predict the mean listener scores are low, yet their combined correlation with the mean listener scores are high; in addition, duration adds significant predictive power to the $F_0$-based dynamic difference measure. This shows that these two measures quasi-independently contribute to stress. This is further anecdotally illustrated in Figure 9, which shows data from three children that received near-identical mean listener scores in the Focus task. As the figure indicates, they differ substantially in terms of the degrees to which they use the three acoustic features. We note that the listeners were asked only to determine which utterance belonged to which category and rate their certainty. It is thus possible that, while these three children expressed the focus contrast with equal *discriminability*, as reflected by the near-equal mean listener scores, they might have differed in *atypicality*. For example, the minimal reliance on the duration feature of two of the three children (who happened to be in the autism group) may be associated with peculiarly sounding speech. Thus, while these data points are intriguing, the issue of whether communicatively effective and typical prosodic contrasts can be expressed with a wide range of balances between the three features, or rather must have a specific balance, needs to be looked at in more detail. Additionally, it is possible that atypicality of prosodic contrasts may reside in these contrasts being expressed too strongly, even when the balance of the features is typical.[3]

**v.** Despite the good results in the Phrasing Task (Figure 6), we are not satisfied with the usage of a single feature – duration. As pointed out earlier, intonational cues appropriate for the various linguistic structures need to be explored for this task.

**vi.** For the Pragmatic Style Task (Figure 7), the robust $F_0$ maximum is a powerful predictor, with $B_4$ adding a non-significant amount of predictive power. Informal

[3]We thank one of the reviewers for making this point.

listening reveals a clear difference in voice quality between the two conditions, and we suspect that $B_4$ does not adequately capture this aspect. Exploratory analyses have not made us confident that we will find reliable, robust dynamic measures (such as those based on temporal structure or on dynamic differences) that can discriminate between the two styles.

**vii.** The data from the affect task (Figure 8) confirm the predictive power of the (b2+b3)-(b1+b4) contrast discovered in the actor data. Also worth noting is the fact that while the correlations between the listeners are much weaker for the per-speaker data than for the per-utterance data, the performance of the automated measures is about the same. It is unknown whether this is due to the per-speaker scoring, in contrast to the per-utterance scoring, taking into account the appropriateness of the responses, or to some other difference between the two types of scores.

## 6. Conclusions

The automated methods proposed in this paper succeeded in approximating human ratings in reliability, as assessed via correlations with auditory-perceptual mean listener ratings. In addition, the objective measures were superior to the conventional method of assessment in which the examiner makes real-time judgments and verifies these offline. These automated methods could be of immediate practical value in terms of substantial labor savings and enhanced reliability.

More important, however, are the principles that underlie the methods and that were spelled out in the Introduction, including the usage of highly specific elicitation methods; a prosodic minimal pairs design; robust acoustic features; capturing both global and dynamic features; and weighted combinations of multiple, maximally independent acoustic features. We are currently developing additional methods based on these same principles for other tasks in our protocol.

Several issues need to be addressed as we move forward. First, we would like to develop methods based on this first generation that detect new markers of neurological disorders – markers that are not audible to the human ear. Such markers include both acoustic features that have a too low SNR to be humanly detectable or are too complex, such as those based on statistical properties of relatively long speech fragments. Second, our methods, while not requiring human judgment, are not fully automatic because they require human labeling and segmentation, some at the word level and others at the phonetic segment level. In theory, automatic segmentation methods may make similar mistakes in segmenting the two utterances making up a prosodic minimal pair (e.g., locating the vowel-nasal boundary too early in both shinaig and shinaig), thereby possibly canceling the effects of the error, but this needs to be demonstrated with actual automatic segmentation systems. Third, the methods need to be extended to unrestricted speech. Key for preserving the prosodic minimal pairs feature of our methods will be the creation of algorithms that detect *quasi prosodic minimal pairs* in broadly elicited speech samples. Quasi prosodic minimal pairs are pairs of prosodically contrastive words or phrases that are phonemically similar but not identical. Finally, we need to develop methods that not only assess whether a speaker can consistently express a certain prosodic contrast but also whether the speaker expresses this contrast with an appropriate balance of prosodic features.

## Acknowledgments

We thank Sue Peppé for making available the pictorial stimuli and for granting permission to us for creating new versions of several PEPS-C tasks; Lawrence Shriberg for helpful comments on an earlier draft of the paper, in particular on the LSR section; Rhea Paul for helpful comments on an earlier draft of the paper and for suggesting the Lexical

Stress and Pragmatic Style Tasks; the clinical staff at CSLU (Beth Langhorst, Rachel Coulston, and Robbyn Sanger Hahn) and at Yale University (Nancy Fredine, Moira Lewis, Allyson Lee) for data collection; senior programmer Jacques de Villiers for the data collection software and data management architecture; Meg Mitchell and Justin Holguin for speech transcription; and the parents and children for participating in the study. This research was supported by a grant from the National Institute on Deafness and Other Communicative Disorders, NIDCD 1R01DC007129-01; a grant from the National Science Foundation, IIS-0205731; by a Student Fellowship from AutismSpeaks to Emily Tucker Prud'hommeaux; and by an Innovative Technology for Autism grant from AutismSpeaks. The views herein are those of the authors and reflect the views neither of the funding agencies nor of any of the individuals acknowledged.

## References

Adami, AG.; Mihaescu, R.; Reynolds, DA.; Godfrey, JJ. Modeling prosodic dynamics for speaker recognition. Proc. ICASSP '03; 2003. IV-06-05

Barlow, RE.; Bartholomew, DJ.; Bremner, JM.; Brunk, HD. Statistical Inference Under Order Restrictions. Wiley; New York: 1972.

Berk S, Doehring DG, Bryans B. Judgments of vocal affect by language-delayed children. Journal of Communication Disorders 1983;16:49–56. [PubMed: 6853751]

Berument S, Rutter M, Lord C, Pickles A, Bailey A. A autism screening questionnaire: diagnostic validity. The British Journal of Psychiatry 1999;175:444–451. [PubMed: 10789276]

Cardozo B, Ritsma R. On the perception of imperfect periodicity. IEEE Trans Audio Electroacoust 1968;16:159–164.

Cohen J. A coefficient of agreement for nominal scale. J Educational Psychological Measurement 1960;20(1):37–46.

Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. Journal of Speech and Hearing Research 1969a;12:246–269. [PubMed: 5808852]

Darley FL, Aronson AE, Brown JR. Clusters of deviant speech dimensions in the dysarthrias. Journal of Speech and Hearing Research 1969b;12:462–496. [PubMed: 5811846]

Dollaghan C, Campbell T. Nonword repetition and child language disorder. Journal of Speech, Language and Hearing Research 1998;41:1136–1146.

DSM-IV-TR. Diagnostic and statistical manual of mental disorders. American Psychiatric Press; 2002.

Ekman, P.; Friesen, W. Pictures of Facial Affect. Consulting Psychologists Press; Palo Alto, CA: 1976.

Entropic Research Laboratory, Inc. ESPS Programs A–L. 1996

Gilberg, C.; Ehlers, S. High functioning people with autism and Asperger syndrome. In: Schopler, E.; Mesibov, G.; Kunce, L., editors. Asperger Syndrome or High Functioning Autism?. Plenum Press; New York: 1998. p. 79-106.

Gotham K, Risi S, Pickles A, Lord C. The Autism Diagnostic Observation Schedule (ADOS): revised algorithms for improved diagnostic validity. J Autism Dev Disord 2007;37(4):613–627. [PubMed: 17180459]

Gotham K, Risi S, Dawson G, Tager-Flusberg H, Joseph R, Carter A, Hepburn S, McMahon W, Rodier P, Hyman S, Sigman M, Rogers S, Landa R, Spence A, Osann K, Flodman P, Volkmar F, Hollander E, Buxbaum J, Pickles A, Lord C. A Replication of the Autism Diagnostic Observation Schedule (ADOS) Revised Algorithms. J Am Acad Child Adolesc Psychiatry 2008;47(6):642–51. [PubMed: 18434924]

Grabe, E.; Post, B. Intonational Variation in English. Proceedings of the Speech Prosody 2002 Conference; 11-13 April 2002; Aix-en-Provence: Laboratoire Parole et Langage; 2002. p. 343-346.

Grabe E, Post B, Nolan F, Farrar K. Pitch accent realization in four varieties of British English. Journal of Phonetics 2000;28(2):161–185.

Hirschberg, J. A Corpus-Based Approach to the Study of Speaking Style. In: Horne, M., editor. Festschrift in Honor of Gosta Bruce. Kluwer; Dordrecht: 2000.

Hirschberg, J. Prosodic and Other Acoustic Cues to Speaking Style in Spontaneous and Read Speech. Proceedings of ICPhS-95; Stockholm. August; 1995. p. 36-43.

House A, Rowe D, Standen P. Affective prosody in the reading voice of stroke patients. J Neurol Neurosurg Psychiatry 1987 July;50(7):910–912. 1987. [PubMed: 3625214]

Elemetrics. Multi-Dimensional Voice Program (MDVP). Pine Brook, NJ: 1993. Computer program

Kent RD. Hearing and Believing: Some Limits to the Auditory-Perceptual Assessment of Speech and Voice Disorders. Am J Sp-Lang Path 1996;5:7–23.

Klabbers, E.; Mishra, T.; van Santen, J. Analysis of affective speech recordings using the superpositional intonation model. Proceedings of the 6th ISCA workshop on speech synthesis (SSW6); Bonn, Germany. 2007. p. 339-344.

Klatt D. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. J Acoust Soc Am 1976;59:1208–1221. [PubMed: 956516]

Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. J Acoust Soc Am 1998;104:1598–1608. [PubMed: 9745743]

Le Dorze G, Ryalls J, Brassard C, Boulange N, Ratté D. A Comparison of the Prosodic Characteristics of the Speech of People with Parkinson's Disease and Friedreich's Ataxia with Neurologically Normal Speakers. Folia Phoniatr Logop 1998;50:1–9. 1998. [PubMed: 9509733]

Lee S, Potamianos A, Narayanan S. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. J Acoust Soc Am 1999;105:1455–1468. [PubMed: 10089598]

Lord C, Risi S, Lambrecht L, Cook E, Leventhal B, DiLavore P, Pickles A, Rutter M. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord 2000;30:205–223. [PubMed: 11055457]

Mackey L, Finn P, Ingham R. Effect of speech dialect on speech naturalness ratings: a systematic replication of Martin, Haroldson, and Triden (1984). J Speech Lang Hear Res 1997 Apr;40(2):349–60. 1997. [PubMed: 9130203]

McCann J, Peppé S. Prosody in autism spectrum disorders: A critical review. International Journal of Language and Communication Disorders 2003;38:325–350. [PubMed: 14578051]

Marasek, K. Glottal correlates of the word stress and the tense/lax opposition in German. Proceedings ICSLP96; 1996. p. 1573-1576.

Miao, Q.; Niu, X.; Klabbers, E.; van Santen, J. Effects of Prosodic Factors on Spectral Balance: Analysis and Synthesis. Prosody; 2006; Dresden, Germany. 2006.

Milroy, L.; Milroy, J. Belfast: change and variation in an urban vernacular. In: Trudgill, P., editor. Sociolinguistic patterns in British English. Edward Arnold; London: 1978. p. 19-37.

Murry T, Doherty E. Selected acoustic characteristics of pathologic and normal speakers. Journal of Speech and Hearing Research 1980;23:361–369. [PubMed: 7442196]

Ofuka, E.; Valbret, H.; Waterman, M.; Campbell, N.; Roach, P. The role of F0 and duration in signalling affect in Japanese: anger, kindness and politeness. ICSLP-1994; 1994. p. 1447-1450.

Ofuka E, McKeown J, Waterman M, Roach P. Prosodic cues for rated politeness in Japanese speech. Speech Communication 2000;32(3):199–217.

Paul R, Augustyn A, Klin A, Volkmar F. Perception and production of prosody by speakers with autism spectrum disorders. Journal of Autism and Developmental Disorders 2005;35:201–220.

Peppé S, McCann J, Gibbon F, O'Hare A, Rutherford M. Assessing prosodic and pragmatic ability in children with high-functioning autism. Journal of Pragmatics 2006;38:1776–1791.

Peppé S, McCann J, Gibbon J, O'Hare A, Rutherford M. Receptive and expressive prosodic ability in children with high-functioning autism. Journal of Speech, Language, and Hearing Research 2007;50:1015–1028.

Peppé S, McCann J. Assessing intonation and prosody in children with atypical language development: The PEPS-C test and the revised version. Clinical Linguistics and Phonetics 2003;17:345–354. [PubMed: 12945610]

Plant G, Öster A. The effects of cochlear implantation on speech production. A case study. STL-QPSR 1986;27(1):65–86.

Roark, R.; Hosom, JP.; Mitchell, M.; Kaye, J. Automatically derived spoken language markers for detecting Mild Cognitive Impairment. Proceedings of the 2nd International Conference on Technology and Aging (ICTA); 2007; 2007.

Scherer KR. Vocal communication of emotion: a review of research paradigms. Speech Commun 2003;40:227–256.

Shriberg, L.; Allen, C.; McSweeny, J.; Wilson, D. PEPPER: Programs to examine phonetic and phonologic evaluation records. Madison: Waisman Research Center Computing Facility, University of Wisconsin—Madison; 2001. Computer software

Shriberg L, Ballard K, Tomblin J, Duffy J, Odell K, Williams C. Speech, prosody, and voice characteristics of a mother and daughter with a 7;13 translocation affecting FOXP2. J Speech Lang Hear Res 2006 Jun;49(3):500–25. [PubMed: 16787893]

Shriberg LD, Campbell TF, Karlsson HB, Brown RL, McSweeny JL, Nadler CJ. A diagnostic marker for childhood apraxia of speech: The lexical stress ratio. Clinical Linguistics and Phonetics 2003;17:549–574. [PubMed: 14608799]

Siegel DJ, Minshew NJ, Goldstein G. Wechsler IQ profiles in diagnosis of high-functioning autism. Journal of Autism and Developmental Disorders 1996;26(4):389–406. [PubMed: 8863091]

Silverman, K.; Kalyanswamy, A.; Silverman, J.; Basson, S.; Yaschin, D. Synthesizer intelligibility in the context of a name and- address information service. Proceedings of Eurospeech '93; September 1993; Berlin, Germany. 1993. p. 2169-2162.

Sluijter A, van Heuven V. Spectral balance as an acoustic correlate of linguistic stress. The Journal of the Acoustical Society of America 1996 October;100(4):2471–2485. [PubMed: 8865652]

Sönmez K, Shriberg E, Heck L, Weintraub M. Modeling dynamic prosodic variation for speaker verification. Proc ICSLP-1998. 1998

The Snack Sound Toolkit. Royal Inst. Technol; 2006. http://www.speech.kth.se/snack/

Sutton, S.; Cole, R., et al. Universal Speech Tools: the CSLU Toolkit. Proceedings of the International Conference on Spoken Language Processing (ICSLP); Sydney, Australia. 1998. p. 3221-3224.

Tucker Prud'hommeaux, E.; van Santen, J.; Paul, R.; Black, L. Automated measurement of expressive prosody in neurodevelopmental disorders. International Meeting For Autism Research; 2008; London, UK. 2008.

Tversky A. Intransitivities of Preferences. Psychological Review 1969;84:327–352.

van Lancker Sidtis D, Pachana N, Cummings JL, Sidtis JJ. Dysprosodic speech following basal ganglia insult: Toward a conceptual framework for the study of the cerebral representation of prosody. Brain and Language 2006;97(2):135–153. [PubMed: 16271755]

van Santen J. Contextual effects on vowel duration. Speech Communication 1992;11:513–546.

van Santen J. Perceptual experiments for diagnostic testing of text-to-speech, systems. Computer Speech & Language 1993;7:49–100.

van Santen J, Hirschberg J. Segmental effects on timing and height of pitch contours. Proc ICSLP 1994;94:719–722.

van Santen, J.; Sproat, R. High-accuracy automatic segmentation. Proceedings of Eurospeech 99; Budapest, Hungary. 1999.

van Santen J, Shih C. Suprasegmental and segmental timing models in Mandarin Chinese and American English. Journal of the Acoustical Society of America 2000;107:1012–1026. [PubMed: 10687710]

van Santen, J.; Möbius, B. A quantitative model of F0 generation and alignment. In: Botinis, A., editor. Intonation - Analysis, Modeling and Technology. Kluwer; Dordrecht: 2000. p. 269-288.

van Santen, J.; Niu, X. Prediction and Synthesis of Prosodic Effects on Spectral Balance of Vowels. Fourth IEEE Workshop on Speech Synthesis; Santa Monica, CA. 2002; 2002.

van Santen J, Klabbers E, Mishra T. Towards measurement of pitch alignment. Italian Journal of Linguistics. Special issue on Autosegmental-metrical approaches to intonation in Europe: tonal targets and anchors 2006;18:161–188.

van Santen, J.; Paul, R.; Black, L.; Tucker, E. Quantitative Analysis of Grammatical and Pragmatic Prosody in Autism Spectrum Disorder. International Meeting For Autism Research; 2007; Seattle, Washington. 2007.

van Santen, J.; Tucker Prud'hommeaux, E.; Paul, R.; Black, L.; Shriberg, L. Expressive prosody in autism: Effects of prosody function and processing demands. International Meeting For Autism Research; 2008; London, UK. 2008.

Wingfield A, Lombardi L, Sokol S. Prosodic features and the intelligibility of accelerated speech: Syntactic versus periodic segmentation. Journal of Speech, and Hearing Research 1984;7:128–134. [PubMed: 6716998]

Zhang Y, Jiang J. Acoustic Analyses of Sustained and Running Voices from Patients with Laryngeal Pathologies. Journal of Voice 2008;22(1):1–9. [PubMed: 16978835]

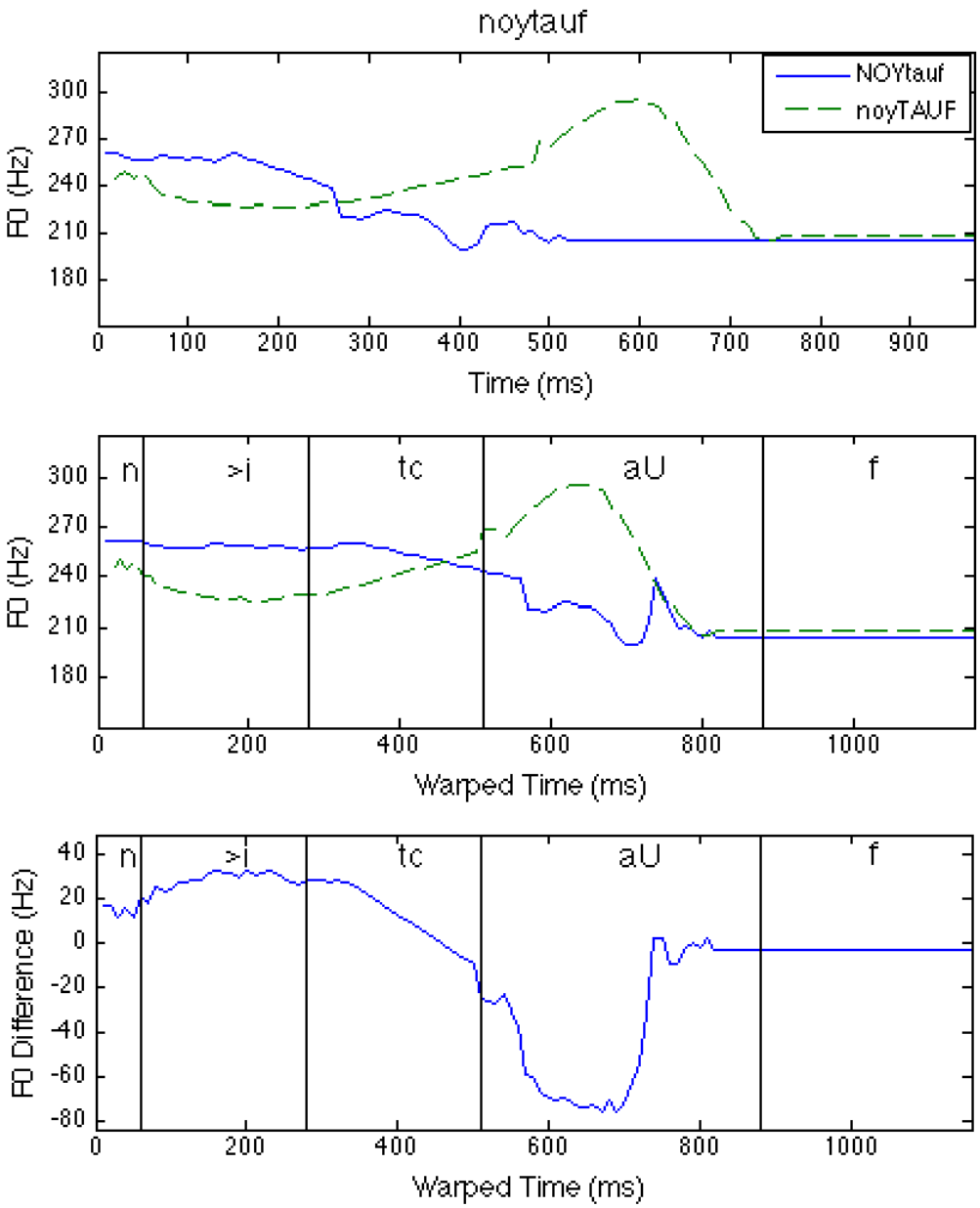


**Figure 1.**
Preparatory steps for Dynamic Difference measures. Horizontal axes: Time. Vertical axes: Frequency (or frequency difference), in Hz. (Top) Original contours of "noytauf" (solid lines) and "noytauf" (dotted line). (Center) Contours after time warping. (Bottom) Difference curve, exhibiting the up-down-up (*udu*) pattern.

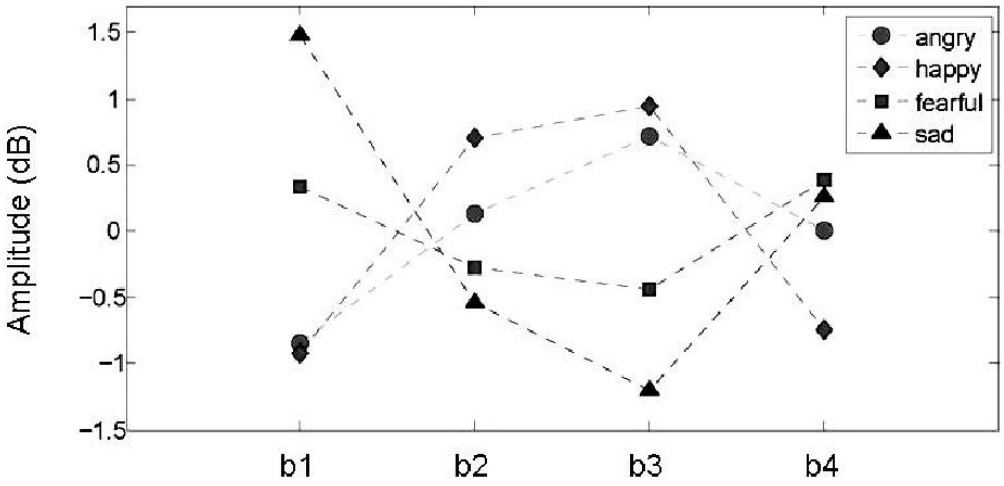


**Figure 2.**
Average values of the four affects on amplitude bands $b_1$-$b_4$ (actor data).

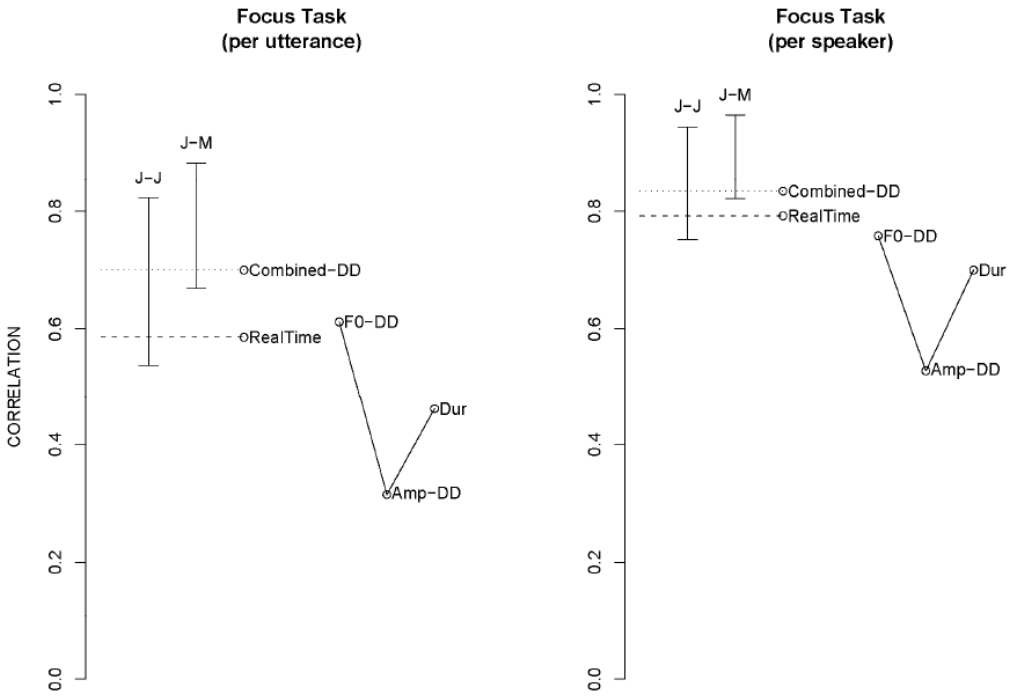


**Figure 3.**

Focus Task. Except for J-J, which indicates the range of inter-judge correlations, all data points indicate correlations with the mean listener scores. J-M indicates the range of judge-mean listener correlations. Combined-DD is the correlation between the mean listener scores and the best prediction of mean listener scores generated by application of (cross-validated) multiple regression to $F_0$-DD (fundamental frequency based *dynamic difference* [DD] measure), Amp-DD (amplitude based dynamic difference measure), and Dur (duration based measure). Finally, RealTime indicated the correlation of the verified real-time scores with the mean listener scores.

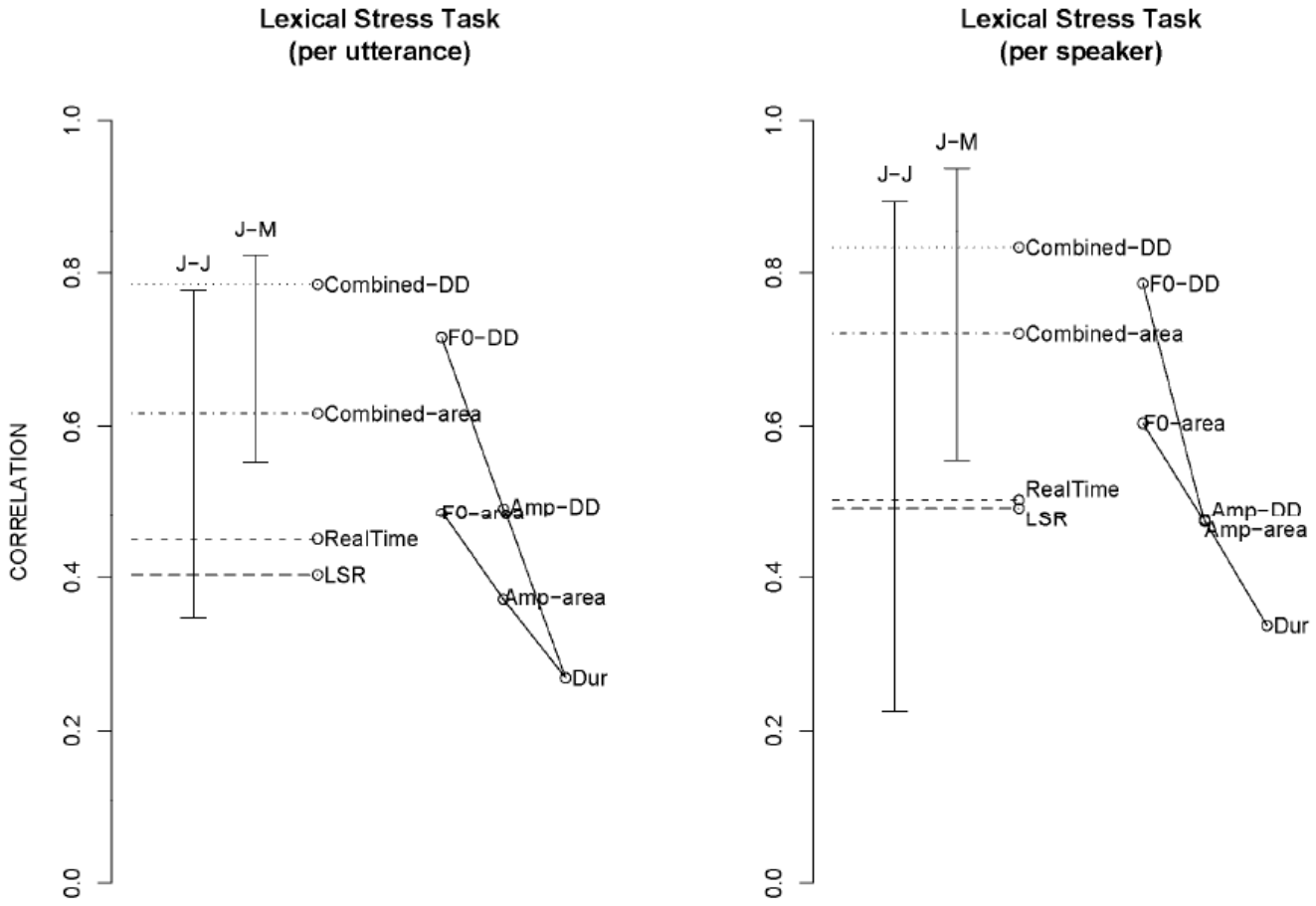


**Figure 4.**

Lexical Stress Task. Conventions are similar to those in Figure 3, with these exceptions. Combined-area is the correlation between the mean listener scores and the best prediction of mean listener scores generated by application of (cross-validated) multiple regression to $F_0$-area (based on fundamental frequency area, i.e., the product of average $F_0$ and vowel duration), Amp-area (based on amplitude area), and Dur (duration based measure). LSR is the Lexical Stress Ratio based measure, using the same weights as in Shriberg et al. (2001).

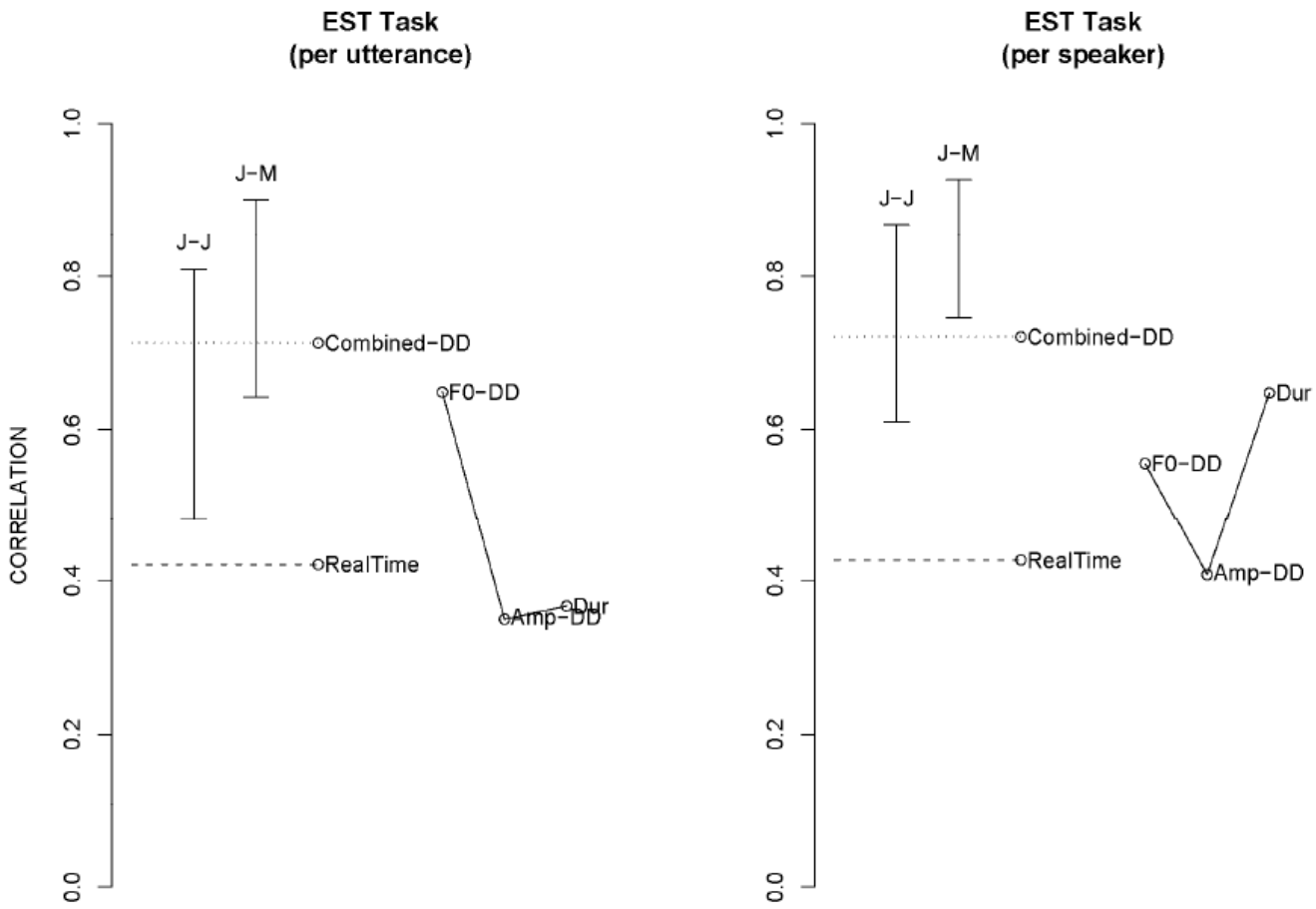


**Figure 5.**
Emphatic Stress Task (EST). Conventions are the same as those in Figure 1.

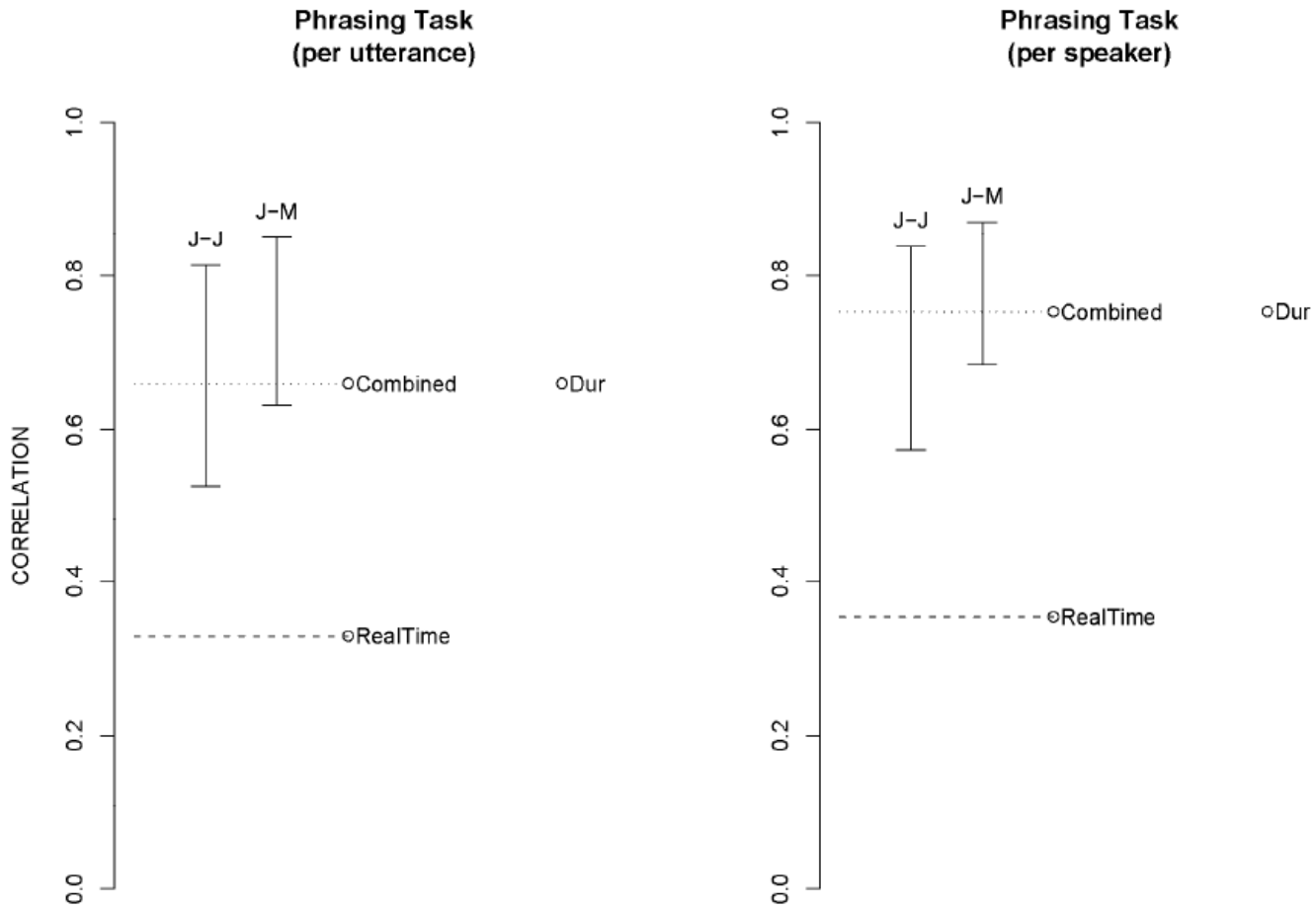


**Figure 6.**
Phrasing Task. Conventions are similar to those in Figure 1. Since only one predictor is used, Dur, the Combined score's correlation with the mean listener scores is the same as that of this predictor.

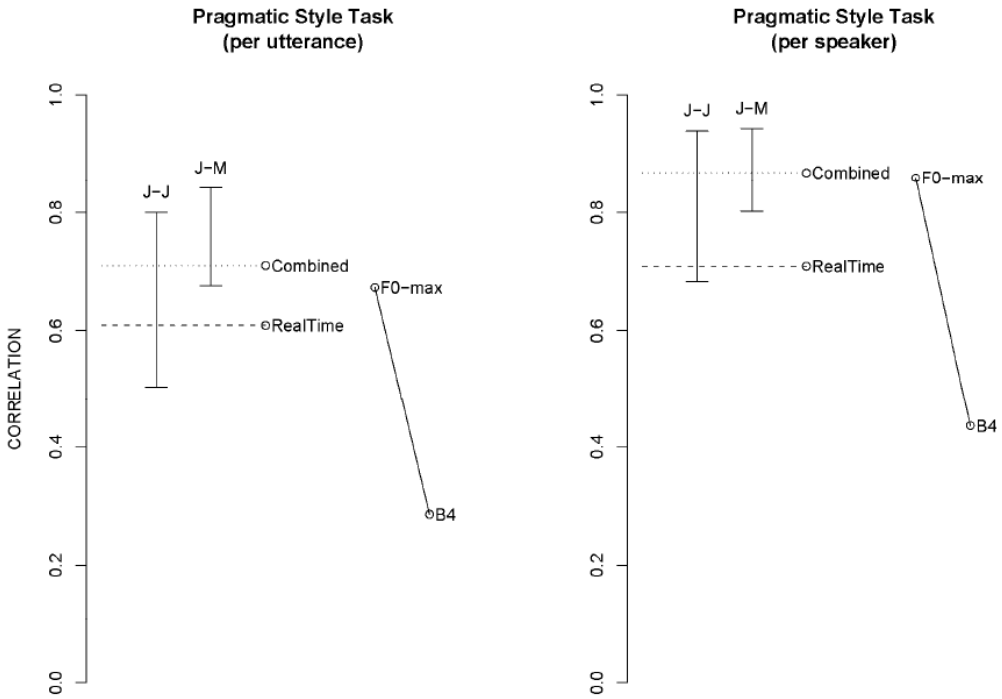


**Figure 7.**
Pragmatic Style Task. Conventions are similar to those in Figure 1. Predictors used are a robust measure of the maximal $F_0$ value and the fourth frequency band, which covers primarily fricative or breathy energy.

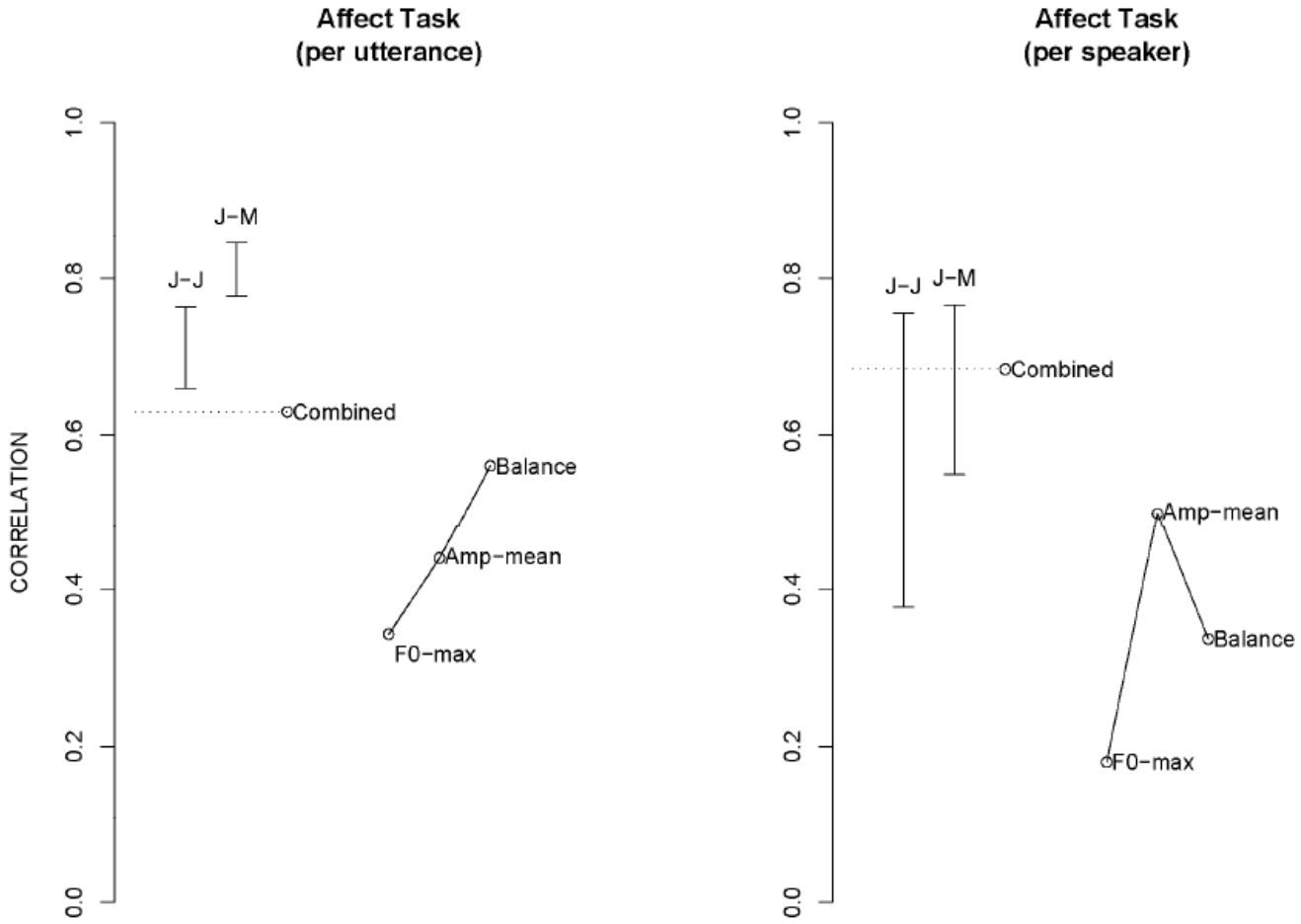


**Figure 8.**

Affect Task. Conventions are the same as in earlier figures, with these exceptions. Amp-mean is the average amplitude of the utterance. Balance is the quantity (B2+B3)-(B1+B4), averaged over the utterance. See text for an explanation of the per-speaker analyses.

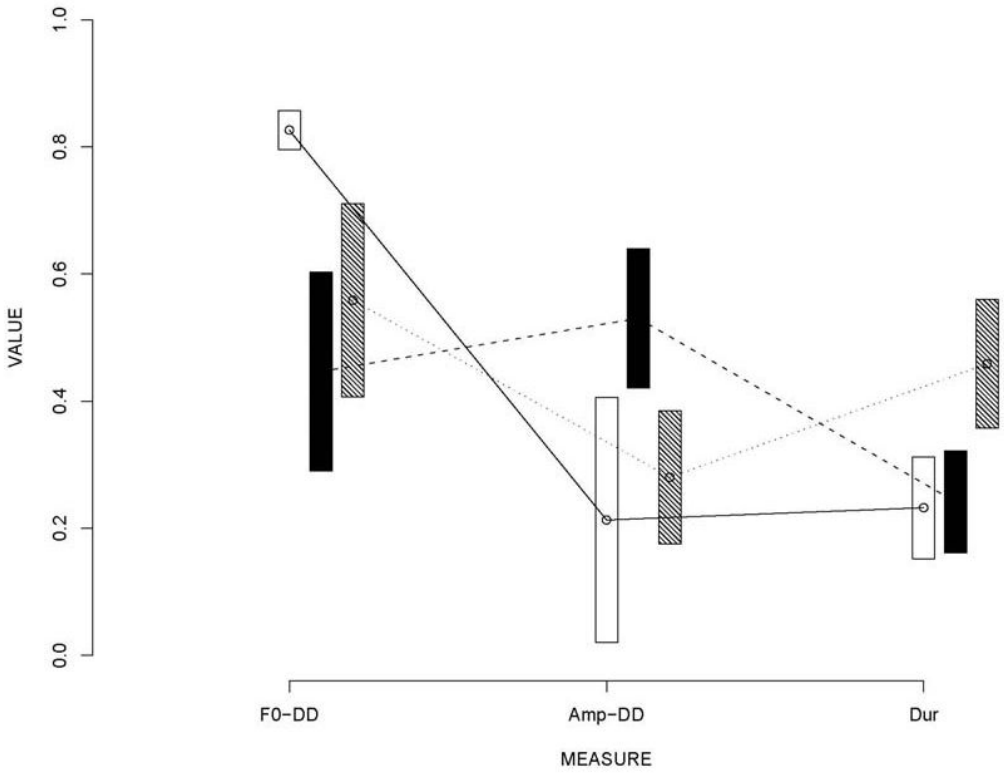


**Figure 9.**

Mean values (with standard errors, based on 8 utterances per child) of $F_0$ dynamic difference, amplitude dynamic difference, and duration based measures for three children having approximately the same mean listener scores (in the 0.80 – 0.83 range) in the Focus Task. The children are identified by the shading of the error bars. The Figure illustrates the compensatory aspect of prosody, with each of the children expressing stress with a different balance between the three prosodic features.