MICROCOMPUTER-BASED INTERACTIVE PROSODY WORKSTATION

George D. Allen and V. Paul Harper

Purdue University and Harper Associates
West Lafayette, IN

ABSTRACT

One of the difficult problems facing teachers of phonetics is the lack of tools for training prosody (i.e., intonation, stress, and syllable rhythm). Our Interactive Prosody Trainer is a low-cost, microcomputer-based system for interactively teaching speech prosody. Based on a digital signal processing chip and an easy-to-use graphical interface, this device does two different, interrelated jobs. On output from the host microcomputer, it synthesizes models of utterances from stored LPC and prosodic control parameters. On input, it extracts the fundamental frequency and intensity of the user's productions, for comparison with the model. Similarities and differences between the two productions are then highlighted for the user.

1. INTRODUCTION

The teaching of prosody has not taken adequate advantage of modern speech technology. Beyond what is found in books or in the heads of phonetics teachers, there are just two audio-cassette resources and some not-very-interactive hardware. One of the audio-cassettes is the demonstration tape accompanying Cruttenden's text [1]. The author has simply read the examples from the text, as illustrations of the points he makes -- hardly a compelling pedagogy. The other cassette material is Allen's auto-instructional tutorial [2], used successfully for several years in his own phonetics classes but not disseminated widely.

Two (very similar) hardware products exist for training prosody, namely the Visipitch (Kay Elemetrics Corp.) and the PM Analyzer (Voice Identification, Inc.). Both devices extract fundamental frequency (F0) and intensity in real time and display them on a computer monitor. Both permit the visual comparison of a student's response to a teacher's model. Unfortunately, neither one can play back the model and/or the response for auditory comparison by the user, and both require the teacher to be present to evaluate the student's response. And they are both expensive. In other words, these devices are helpful aids for the teaching of prosody, but they require extensive one-to-one interaction with a trained professional, and a majority of training programs can afford to buy at most one.

Interestingly enough, Lane & Buiten [3] showed over 25 years ago that an interactive computer workstation could teach prosody effectively. Their so-called "Speech Auto-Instructional Device" (SAID) required users to match, as closely as possible, either the F0, the intensity, or the syllable timing of a model utterance. Using analog F0 and intensity extractors, plus a DEC PDP/1 minicomputer to calculate the match between model and response, the device cycled users successively among the three prosodic features until all three had converged to an acceptable degree. As successful as the SAID was in training fluent prosody, it is perhaps surprising that its principles have never been extended to modern microcomputers and digital signal processing technology. That time has now come.

2. A SAMPLE SESSION

The user, a young Chinese student who is improving his English, sits in front of the display screen and presses the **Model** button on his keypad. The computer presents the phrase "Good morning" with mid-level pitch on "Good" and a high-falling pitch on "morning." As the utterance is played out, white dots follow along the fundamental frequency (F0) and intensity traces on the screen. If he wishes, he can press the **Model** button to hear the utterance again.

When he is ready to respond, the user presses and holds down the **Talk** key. As he mimics the model with reiterant speech, which consists of repeated /ma/ syllables [4], his F0 and intensity are drawn on the screen in a different color and width of line from the model. Differences between the model and response prosodies are highlighted with shading, and "scores" are generated comparing the F0 and intensity traces and the timing of the syllables.

By pressing either the **Model** key or the **Response** key, the student can then hear "Good morning" uttered with the original prosody or with the prosody of his own response. As the utterance is played out, white dots trace out the prosody curves of the utterance to which he is listening.

At this point, the student has several choices. He may wish to produce another response, which he does by pressing and holding down the **Talk** key. Or, he may wish to review an earlier response, which he does by pressing a left-pointing arrow key the appropriate number of times (once for the previous response, etc.). Data from earlier responses are then presented on the screen, and he can again listen to either the model or the response. Finally, he may wish to move on to another utterance, which he does by pressing the **New** key.

As he works, all of his data are saved in a file, which is then written to disk at the session's completion. Later, the instructor will review this file, evaluating the student's progress.

3. WORKSTATION FEATURES

A prototype has been built using off-the-shelf components and well-documented speech processing algorithms. The hardware consists of a personal computer (PC) with a VGA color graphics display, microphone, speaker, and an audio I/O card with a digital signal processing (DSP) chip. The software consists of F0 and intensity extraction, speech encoding, and human interface modules.

Model utterances are stored in the PC, using linear prediction coefficients (LPC) plus prosodic information. There are two important advantages to using LPC coded models. One is simply the reduction in storage demands, in comparison with digitized speech. The other advantage is more important, however. Effective use of a prosody trainer requires that the model text be presented with either the model or the user's prosody. Since users of our device will respond with reiterant speech, the extracted F0 and intensity data can be substituted for the model data and used to drive the LPC re-synthesis. Thus, the user can easily toggle back and forth between the model as originally presented and the same words uttered with the prosody of the user's response.

Many algorithms exist for extracting F0 from speech [5]. Because reiterant speech is fully and continuously voiced, most F0 extraction algorithms work (equally) well. Speech intensity is usually also obtained as a by-product of this F0 extraction process. Thus, while the user is talking, current estimates of F0 and intensity can be delivered to the PC at the same rate as the model. Before the data are drawn on the screen, they are smoothed, scaled, and sometimes time-normalized. We say "sometimes," because there are situations in which the timing should be corrected, and others in which it should not.

F0 and intensity must also be appropriately scaled, both for the display and for the calculation of the fit between model and response. Both F0 and intensity are scaled logarithmically, to match our perceptions of these acoustic features. Scores for the degree of match between model and response are then calculated as weighted averages of the differences between the data values. The use of weights permits focusing of attention on important speech features, such as vowel nuclei, without corruption by consonantal gestures.

Adjusting the timing of the response is another case of a difficult task made easier by the use of reiterant speech. Consonant/vowel (C/V) boundaries are relatively easy to locate from the intensity trace of repeated /ma/ syllables. These boundaries can then be aligned successively with pre-stored C/V boundaries of the model. Data points are deleted or interpolated in the F0 and intensity traces, as required, to match up the timing for each segment of the utterance. Finally, a score for the accuracy of timing can be generated as a weighted sum of the adjustments required to align the response to the model. Again, weighting these sums permits the user to focus on important temporal aspects of the model without becoming distracted by irrelevant features that happen to have been mis-timed.

Since the target users are persons with little or no experience in computer use, the interface must be simple and easily learned. As described in §2,above, a session with the Interactive Prosody Trainer requires the user to listen to a spoken phrase, mimic it using reiterant speech, and then view a graphical display showing his response compared to the original. He is then able to listen to the model phrase with its original prosody or with the prosody of his own response. This is similar to the SAID procedure [1], referred to earlier, in which learners alternately mimicked the F0, intensity, or timing of model phrases. In addition, however, our device permits the user to listen to the target phrase with his or her own prosody, in direct comparison with the model. This form of feedback is crucial for successful prosodic training.

4. REFERENCES

[1] CRUTTENDEN, A. (1986) *Intonation.* New York: Cambridge University Press.

[2] ALLEN, G. D. (1984) *Transcribing the Prosodic (Suprasegmental) Features of English.* Short Course presented at the Annual Meeting of the American Speech-Language-Hearing Association, San Francisco, CA, November 17, 1984.

[3] LANE, H. L. & BUITEN, R. (1965) A self-instructional device for conditioning accurate prosody. *International Review of Applied Linguistics*, 3, 205-219.

[4] NAKATANI, L. H. & SCHAFFER, J. A. (1978) Hearing "words" without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America*, 63, 234-245.

[5] HESS, W. (1983) *Pitch Determination of Speech Signals.* New York: Springer Verlag.