A Platform for Automated Acoustic Analysis for Assistive Technology

Suzanne Boyce
Department of Communication Sciences and Disorders
University of Cincinnati
Cincinnati, Ohio, 45267, USA
boycese@ucmail.uc.edu

Harriet Fell
College of Computer and Information Science
Northeastern University
Boston, Massachusetts, 02115, USA
fell@ccs.neu.edu

Joel MacAuslan
Speech Technology and Applied Research
54 Middlesex Turnpike
Bedford, Massachusetts, , USA
joelm@staranalyticalservices.com

Lorin Wilde
Boston University

Boston, Massachusetts, 02180, USA
wildercom@gmail.com

Abstract

The use of speech production data has been limited by a steep learning curve and the need for laborious hand measurement. We are building a tool set that provides summary statistics for measures designed by clinicians to screen, diagnose or provide training to assistive technology users. This will be achieved by extending an existing shareware software platform with "plug-ins" that perform specific measures and report results to the user. The common underlying basis for this tool set is a Stevens' paradigm of landmarks, points in an utterance around which information about articulatory events can be extracted.

1 Introduction

To date, the use of speech production data has been limited by a steep learning curve and the need for laborious hand measurement. Many speech-related studies result in voluminous acoustic data. Many clinicians who design and use assistive technology would like to incorporate acoustic analysis, but have been discouraged because of these technical challenges. We are in the process of developing a set of tools that considerably streamlines the process of analyzing speech production details.

We are building a tool set to provide summary statistics for measures designed by clinicians to screen, diagnose or provide training to patients. This will be achieved by extending an existing shareware software platform with "plug-ins" that perform specific measures and report results to the user. At present, our goal is to use the existing shareware software tool Wavesurfer (Wavesurfer, 2005). The new modules will be set up to report data from a single audio file, or groups of audio files in a standard table format, for easy input to statistical or other analysis software. For example, the data may be imported into a program that correlates speech data with scalp electrode and medication data.

Our tool will include alternative and independently tested algorithms for clinically relevant measures, as well as guidance as to what the speech data may mean.

The common underlying basis for this tool set is a focused set of landmarks derived from Stevens' Lexical Access from Features (LAFF) paradigm (Stevens, 1992, 2002; Liu, 1995; Slifka et al., 2004). In this approach, landmarks are points in an utterance around which information about articulatory events can be extracted.

In what follows, we will describe (1) the theoretical rationale of landmarks, (2) the general utility of landmark processing and several examples of clinically related measures, and (3) our current work on developing tools to make landmark analysis more widely available.

2 Landmarks reflect articulation

Landmark analysis is based on the fact that different sounds produce different patterns of abrupt changes in the acoustic signal simultaneously across wide frequency ranges. For instance, the abrupt increase in amplitude for a broad range of frequencies above 3 kHz can be used to indicate the onset of bursts. Likewise, an abrupt decrease in the same frequency bands can be used to indicate the end of frication. The use of onset and offset data in other frequency bands can be used to indicate sonorancy; i.e., intervals when the oral cavity is relatively unconstricted. Examples based on Liu [1995] are listed below.

g(lottis): marks the onset (+g) or offset (-g) of voicing.

s(yllabicity): marks the onset (+s) or offset (-s) of syllabicity, i.e. onsets and releases of voiced sonorant consonants such as /l/ or /r/, vocal tract closures due to voiced stop consonants such as /b/ or /d/.

b(urst): marks the onset (+b) of the burst of air following stop or affricate consonant release, or the onset of frication noise for fricative consonants. Offsets (-b) mark points where aspiration or frication noise ends abruptly due to a stop closure.

V(owel): marks points of peak amplitude in a sonorant region—that is, a region where voicing is evident [Howitt, 2000].

Although much of the past work using landmark processing has been focused on employing a wide variety of landmarks to recognize the lexical content of speech [Juneja and Espy-Wilson 2003, Slifka, *et al.* 2004], the power of these measures is even more apparent when applied to non-lexical attributes.

3 Applications of Landmark Analysis to Assistive Technology

3.1 Tracking Articulatory Precision

Measuring articulatory precision is important to evaluating efficacy of a treatment or in monitoring disease progression, e.g. in Parkinson's disease.

Given that landmarks reflect articulation, tools based on landmarks can measure and monitor articulatory precision. This can be tracked by setting empirically derived thresholds for the detection of abrupt acoustic changes in specified frequency bands. Recall that changes in the acoustic signal occur simultaneously across wide frequency ranges. When the onset of energy does not exceed threshold in a particular frequency band, i.e., not quite abrupt enough to trigger the detection of a landmark, then no landmark may be assigned. However, since different sounds produce different patterns, changes detected in other bands at that point in time are either a) assigned to a different landmark, or b) considered to be extraneous. Thus, small acoustic differences in the way speech is produced can be tracked as different patterns of landmarks.

In addition to requirements that a tool for general clinical use must be fast and robust, it must be able to handle a wide variety of speaking styles, dialects, and voices. By focusing on landmarks that specify syllable structure and broad phoneme classes, distinctive differences between phonemes can be ignored. Therefore, the tool is less likely to break down due to problems recognizing specific vocabulary while remaining sensitive to changes in the acoustic signal that reflect articulatory precision of speech.

3.2 Evaluating phonological complexity

Development of speech in early infancy includes the ability to produce increasingly complex phonological structure. Patterns of syllable structure in speech output can be tracked using landmarks, again without reference to specific phonemes or words. In Fell *et al.* [2002], landmarks were grouped into standard syllable patterns and syllables were grouped into utterances. Statistics based on these patterns were then reported to

the clinician for various uses in training, screening or diagnosis. Patterns of syllable complexity were used to compute a "vocalization age." This was used in turn to derive screening rules that clinically distinguish infants who may be at risk for later communication or other developmental problems from typically developing infants.

3.3 Measuring and Evaluating "Clear Speech"

"Clear Speech" is an intelligibility-enhancing style of speech that is used to improve communication outcomes. Listeners with hearing impairment derive significant benefit from being addressed with clearly articulated speech. Speech that is more clearly articulated contains more abrupt acoustic changes. The result is that speech with different levels of intelligibility shows different numbers and combinations of landmarks [Boyce *et al.* 2005, 2007].

3.4 Other Applications

In the UCARE project [1995], Cress reported analyzing 40 hours of pre-existing [2005] videotaped sessions of children with physical or neurological impairments using landmark-based tools.

Fell et al. [2004] reported using landmark analysis to follow the progress of several children with severe speech delays. In this project, 10-minute, in-home audio recordings were processed in real-time on a 2002-era PC laptop.

Wade and Möbius [2007] used automated landmark analysis to study speaking rate effects as a measure of disease progression in Parkinson's disease.

DiCicco and Patel [2008] used automatic landmark analysis on dysarthric speech. This study provides quantitative support for the hypothesis [Deller 1991] that dysarthric speech includes erroneous additional acoustic cues, not only malformed or missing ones.

4 Potential Benefits of Landmark Applications

In a small study, Warner-Czyz and Davis [2010] compared consonant–vowel syllable accuracy in early words of children with normal hearing and children with hearing loss who received cochlear implantation. They found and evaluated, via manual coding, approximately 4000 syllables from 48 hours of recordings. This is a project where automatic landmark analysis might have greatly reduced the effort.

Similarly, in a study on tongue-twisters, Matthew Goldrick (Northwestern University) collected 100 hours of data comprising 20,000 tokens in less than three weeks, but found that it required another 600 hours merely to segment and label the data for further analysis. In personal correspondence about another study on single words, he stated:

> A major 'choke point' for speech production research is the need to manually analyze speech data. Given that many thousands of data points are typically required to gain accurate estimates of probability density functions along phonetic dimensions, hundreds of person-hours are typically required to analyze data from a single simple experiment…. If we could gain access to reliable, highly accurate automated tools, we could change the speed of research by an order of magnitude.

Researchers who currently want to use speech analysis as a tool must accept long periods of hand measurements. This discourages researchers who may be more interested in a particular neurological disease or process than in speech research *per se*. It is notoriously difficult to quantify projects not undertaken, or papers not written, but it is telling that, although each of the studies cited above reported positive results from a study of speech articulation, they exist as relative islands in their respective disciplines. We contend that this situation exists largely because of barriers to entry; that is, we believe that many scientists would like to use speech assessment as part of their research, but elect not to do for lack of a convenient tool. The existence of a convenient tool to detect, measure and track subtle changes in speech articulation would constitute an enabling technology.

5 Tools

5.1 Description

In our own work, we have developed an automatic tool for detecting, counting and analyzing acoustic events in the speech signal that are commonly used by scientists to measure differences in speech articulation.

We are now integrating our system with Wavesurfer for certain researchers (linguists, speech-language pathologists, certain engineering and cognitive-science researchers) with a primary interest in inspecting and interpreting the articulation-related features in the waveforms of a corpus: e.g., the placement of landmarks of each type, patterns of clustering, or identification of non-speech sounds to be excised. (See Figure 1).)

For this version, we are implementing user controls ("widgets") to produce automated measures or types of analyses for speech research such as:

- Voice-onset time, VOT.
- Detection of non-harmonic (and harmonic) voicing.
- Identification and suppression or removal of stray sounds, i.e., non-speech.
- Grouping of landmarks into syllable-like clusters.

(Note that Wavesurfer already provides a general pitch-tracking capability for harmonic voicing.)

The Wavesurfer plug-in will also allow the user to output information about an audio file or a directory of audio files, e.g. all the recordings of a child. This information will be in a tab-delimited text file or a spreadsheet. This will allow the speech scientist

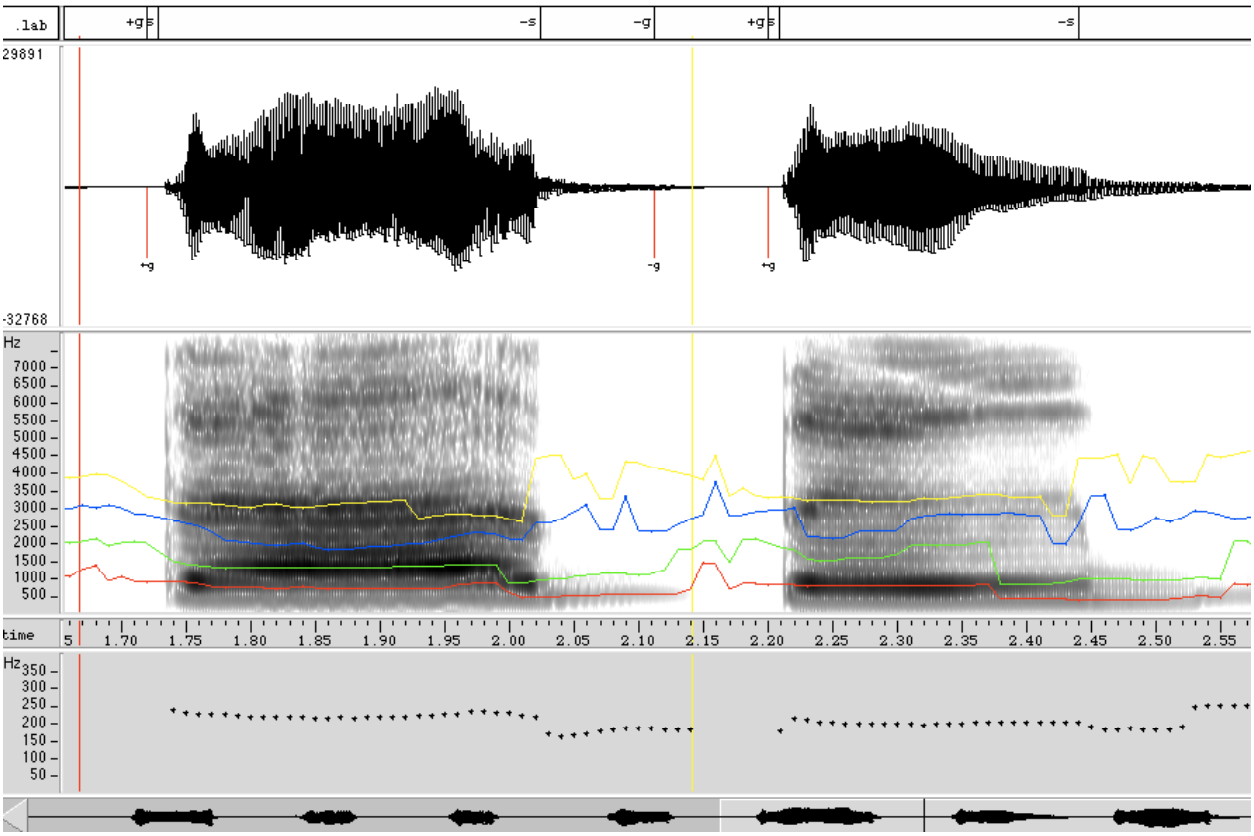

Figure 1: Wavesurfer with landmarks/waveform pane filtered to show only +/-g landmarks, and transcription pane (top) with +/-g and +/-s landmarks

This information will be in a tab-delimited text file or a spreadsheet. This will allow the speech scientist to analyze the output and, for example, to summarize and compare the typically developing children to those diagnosed with autism.

5.2 User Testing

We are currently recruiting potential users to test the system including graduate students and senior researchers in neurosciences and speech-related sciences. So that these users can test the system on a realistic problem, we will provide them with a corpus of annotated, de-identified recordings of children with and without a diagnosis of autism. This will provide context for specific training tasks that we ask of the users and enable them to formulate their own appropriate, if small, research questions that the system can help to answer. We will probe their experiences by logging the questions they have about the system, watching their actions as they attempt to answer the research questions, and asking their opinions of the experience afterward.

6 Requested Features

In an early trial of our Wavesurfer plug-in, a user requested the VOT (voice-onset time) measure. In response to this request, we are now adding a VOT-transcription pane to display the automatically computed voice onset times aligned with the waveform, spectrogram, and displayed information. The information in this pane is also automatically saved to a text file that can be analyzed with other software.

This request also led us to include a popup window to show the vowel-space in a recording. Vowel-space measures are conventionally labor-intensive, thus limited to a few instances of specific vowels, and require that the researcher first identify specific instances of these vowels. On the other hand, vocalic landmarks identify the instants where formant frequencies may be reliably estimated, so our tools can quickly and automatically evaluate the full vowel space of a passage. (See Figure 2.)

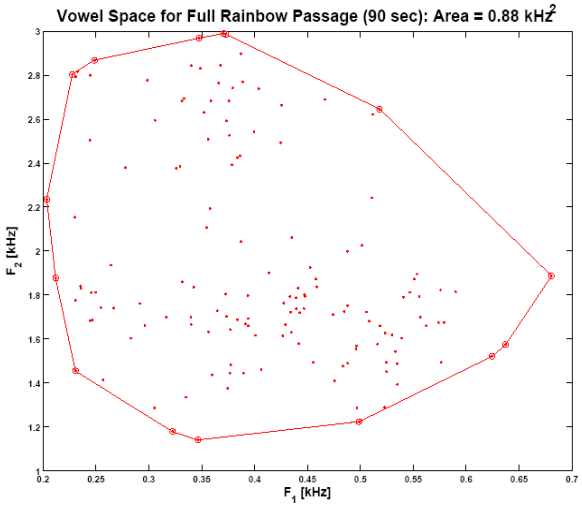

Figure 2: Automatic Vowel-Space Evaluation. Computing the resonant frequencies (formants) at vowel landmarks allows plotting the vowel space, i.e., the scatter of the first two formants against each other. In this case, a female read the complete Rainbow Passage (a standard passage of 3 paragraphs, approx. 90 sec of reading). The system automatically identified all the consonantal and vocalic landmarks, evaluated the formants at ~ 140 stressed vowels, and computed the convex hull ("rubber-band") area, 0.88 kHz^2. Total computation time on a commodity 3 GHz PC was 143 sec (and is directly proportional to the duration of the passage).

7 Challenges for Software development, Challenges for availability

Our algorithms are implemented in MATLAB. Though toolkits that run in MATLAB might be available free, or for a modest price, the MATLAB platform itself is costly, especially for non-academic users. On the other hand, shareware or freeware may have minimal documentation; support that depends entirely on the presence (or absence!) of a knowledgeable user community; and variable standards for testing, correctness, and performance.

A critical hidden cost for any system is the learning curve. For those systems with little documentation and training, this can dwarf the overt costs. Our goal is to make learning easier by creating landmark-processing plug-ins that people can use within software that they already employ.

Such a plan requires a careful balance between the flexibility of a general, extensible system and the simplicity of a small, fixed set of easily documented plug-in capabilities. Our project therefore includes both a small set of simple functions, such as VOT, and software design centered on the needs identified by users from the appropriate research communities. Our design relies on an iterative process of structured interviews and web-based surveys, combined with observations of user experiences with our plug-ins.

This user study extends beyond the matter of functionality and documentation. It also addresses the expectations or requirements for convenient availability, training, and support, and the costs that these imply.

8 Future work

8.1 R – statistical analysis system

We will integrate our software with R (http://www.r-project.org/) for those with a primary interest instead in the derived articulatory-precision information: e.g., syllable production rate, fraction of syllables of a given complexity, or range of vowels.

For this platform, we will implement further user-level functions, with corresponding graphical user interfaces as appropriate, to produce:

- Number of landmarks, optionally excluding those that are automatically detected as noise-related.
- Syllable complexity and statistics of same.
- Utterance complexity.
- Syllable production rate.
- Articulatory precision.
- Vowel space measures.

8.2 Other Platforms

We plan to expand our work to include plugins or packages for integration with a wider (and more powerful) collection of research tools, for example PRAAT, CSL, or even Excel.

8.3 Other Features

We are soliciting input from user communities about the features they would like to see in these tools.

Acknowledgments

This work was funded in part by NIH grant R43 DC010104.

References

Suzanne Boyce, Joel MacAuslan, Ann Bradlow, and Rajka Smiljanič. 2007. Automatic Detection of Differences Between Clear & Conversational Speech, poster presented at *American Speech-Language-Hearing Convention*.

Suzanne Boyce, Ann Bradlow, and Joel MacAuslan. 2005. Landmark analysis of clear and conversational speaking styles, *150th meeting of the Acoustical Society of America*.

Thomas DiCicco and Rupal Patel. 2008. Automatic Landmark Analysis of Dysarthric Speech, *Journal of Medical Speech-Language Pathology*, 16(4):213-219.

Cynthia J. Cress, S. Unrein, A. Weber, S. Krings, H. Fell, J. MacAuslan, and J. Gong. 2005. Vocal Development Patterns in Children at Risk for Being Non-speaking. *ASHA 2005*.

Cynthia J. Cress. 1995. Communicative and symbolic precursors of AAC, Unpublished NIH CIDA Grant: University of Nebraska-Lincoln.

Jack R. Deller, D. Hsu, and Linda J. Ferrier. 1991. On the Use of Hidden Markov Modeling for Recognition of Dysarthric Speech, *Computer Methods and Programs in Biomedicine.* (35)2:125-139.

Harriet J. Fell, Joel MacAuslan, Linda J. Ferrier, Susan G. Worst, and Karen Chenausky. 2002. Vocalization Age as a Clinical Tool. *Procroceedings of the International Conference on Speech and Language Processing*.

Harriet J. Fell, Joel MacAuslan, Cynthia. Cress, Linda J. Ferrier. 2004. visiBabble for Reinforcement of Early Vocalization, *Proceedings of ASSETS 2004*. 161-168.

Wilson Howitt. 2000. *Unpublished Ph.D. dissertation, Massachusetts Institute of Technology*.

Amit Juneja and Carol Espy-Wilson. 2003, Speech Segmentation Using Probabilistic Phonetic Feature Hierarchy and Support Vector Machines. *Proceedings of the International Joint Conference on Neural Networks.*

Sharlene A. Liu. 1995. Landmark Detection for Distinctive Feature-Hyphen Based Speech Recognition, *M.I.T. Doctoral Thesis.*

R, http://www.r-project.org/

Janet Slifka, Kenneth N. Stevens, Sharon Manuel, and Stefanie Shattuck-Hufnagel. 2004. A Landmark-Based Model of Speech Perception: History and Recent Developments. *From Sound to Sense,* 85-90.

Kenneth N. Stevens, 2000. *Acoustic Phonetics*, The MIT Press, Cambridge, Massachusetts.

Kenneth N. Stevens. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features, *Journal of the Acoustic Society of Am*erica. 111(4):1872-1891.

Kenneth N. Stevens, Sharon Manuel, Stefanie Shattuck-Hufnagel, and Sharlene Liu. 1992. Implementation of a model for lexical access based on features*, Proceedings ICSLP (Int. Conf. on Speech & Language Processing)*. 499-502.

Travis Wade, Bernd Möbius. 2007. Speaking rate effects in a landmark-based phonetic exemplar model, *Interspeech 2007*. 402-405.

Wavesurfer.2005. http://www.speech.kth.se/wavesurfer/