

A computer-based course in spectrogram reading

Tim Carmell, John-Paul Hosom

Center for Spoken Language Understanding, OGI, 20000 NW Walker Road, Beaverton, OR 97006.
<http://cslu.cse.ogi.edu>

Ron Cole

Center for Spoken Language Understanding, University of Colorado, Boulder <http://cslu.colorado.edu>

Abstract

We describe a computer-based course in spectrogram reading built with the CSLU Toolkit. The principal teaching tool is SpeechView, a Toolkit component that displays speech waveform files, spectrograms, and labels. These display types are essential for online spectrogram reading. Other Toolkit modules that are used include: (1) an animated face that allows students to view articulator movements; (2) a speech synthesis system that enables students to compare human and synthesized speech; and (3) a phonetic alignment package that performs automatic labeling of speech.

The authors have used these software components to teach three computer-based classes. We have also developed a complete laboratory curriculum, including a speech corpus designed for beginning spectrogram reading. The curriculum and corpus are both summarized in this article.

We conclude by describing our work toward an entirely computer-based course in spectrogram reading, including: (1) enhanced spectrogram displays, (2) new labeling tools, and (3) online textual and graphic materials.

1. Introduction

A course in spectrogram reading has been offered frequently during the past ten years at OGI. Faculty and students concur that this course provides an excellent introduction to speech science. Prior to 1996 the courses used spectrograms and other materials assembled by Victor Zue and his students at MIT [1]. This paper-based mode of teaching has limitations which motivated us to develop a computer-based course in which students can (1) record and view their own speech, (2) access speech data from the many available speech corpora, and (3) modify signal processing and display parameters to achieve optimal readability. The course is being developed in two stages. In the first stage, we have continued to depend on a human teacher, but all of the speech data for the course are available for classroom use. In the second stage, an entire course, including speech data, viewing tools, course text and graphics, and interactive exercises, will be available as a component of the CSLU Toolkit. As of this writing, we have achieved

our first-stage goal through progressive refinement of computer-based tools, speech data and learning exercises.

In section 2, we describe the computer-based software resources and illustrate how the interactive learning tools are used. In section 3, we describe the course curriculum and speech corpus. In section 4, we discuss our plans to distribute a complete course in spectrogram reading as part of a future release of the CSLU Toolkit. A prototype is available via the Internet [2].

2. Course software

2.1 SpeechView

SpeechView is a component of the CSLU Toolkit [3] that enables users to record, play, display and edit acoustic waveforms. Several spectrogram formats with user-defined signal processing and display options are available. SpeechView also allows users to create, display, and edit label files that are time-aligned to the waveforms being labeled. Waveform segments corresponding to a phoneme or word can be played back in isolation from adjacent segments. One or more independent waveform windows, each with zero or more spectrogram and label windows, may be displayed simultaneously within SpeechView. Figure 1 on the next page shows a typical SpeechView display.

The user interface and high-level logic for SpeechView are programmed in Tcl/Tk. The Tcl package mechanism provides access to underlying C language modules for audio, signal processing, image display, and file I/O. This open architecture makes it easy for knowledgeable users to extend the functionality of SpeechView to meet their own needs; BaldiSync is one such extension.

2.2 BaldiSync

The BaldiSync application combines the playback of speech with visible articulator movements. BaldiSync integrates SpeechView's waveform, spectrogram, and label displays with (a) the 3D talking face called Baldi [4], (b) the Festival speech synthesis server [5], and (c) the Toolkit forced alignment package. To synchronize

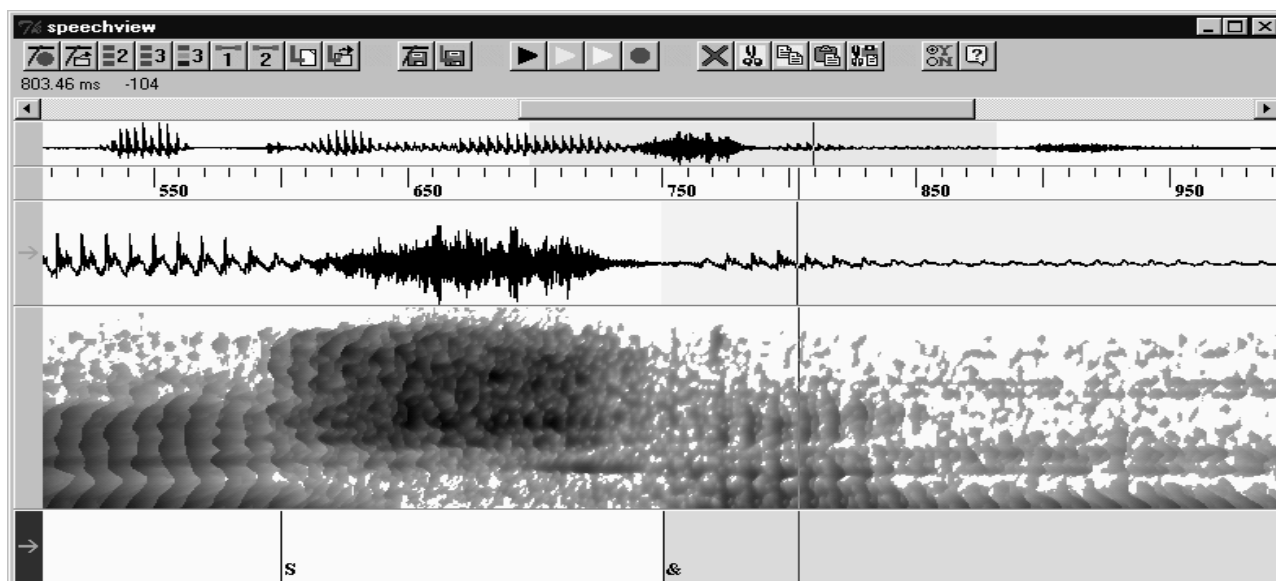


Figure 1 - SpeechView waveform, spectrogram, and label windows

recorded speech and Baldi's facial movements, users record an utterance or read in a waveform file, then supply the text of the utterance as a sequence of English words. BaldiSync assigns a phonetic sequence to each word using a pronunciation dictionary or TTS letter-to-sound rules, and aligns the phonetic segments to the speech waveform. The user can play any portion of the utterance while watching Baldi's movements. Baldi can be made semi-transparent so the tongue is viewed in relation to the teeth, gums, palate and lips from a variety of perspectives. Alternatively, BaldiSync can create and play synthetic speech in synchrony with the facial animation.

3. Laboratory curriculum

The overall goal of the course is to provide a strong foundation for further work in speech science. To attain this goal, the course employs a good textbook [6], lectures, readings, and laboratory work. In this article we focus on the laboratory curriculum. For each unit in our curriculum, the software tools described in the previous section permit students to view relevant spectrogram patterns along with the source waveform, labeled speech segments, and facial articulatory gestures; and to listen to synchronized human or synthetic speech.

The goal of the laboratory sessions is to enable students to recognize the major acoustic patterns in American English speech. These patterns span various time scales and include glottal pulses, formants, bursts, friction and aspiration regions, phonemes, syllables, glottalization, and male versus female speech. Students simultaneously learn (1) to recognize the patterns in spectrograms, (2) to describe the patterns with the appropriate acoustic and phonetic terms, and (3) to explain and theorize about patterns of speech with unifying concepts (e.g., coarticulation) and models (e.g., the source-filter model). This multilevel integration of knowledge permits students to map the patterns to articulatory gestures and phonetic

features evolving in time. The concrete goal is for students to achieve 80% phoneme segmentation and labeling accuracy on clean, short, well-articulated waveforms, with two choices for each segment label. This goal approaches the performance of expert spectrogram readers. [7]

The laboratory begins with the online course entitled *Speech Production and Perception I* from Sensimetrics Corporation [8]. This computer-based course covers the basics of waveforms, spectra, spectrograms, formants, and acoustic phonetics. Students also use SpeechView and BaldiSync to record, play, display, and label waveforms as a prerequisite for further laboratory work. The course laboratory proper begins after these preliminaries.

The laboratory curriculum includes a corpus of speech data available via the Internet [2]; it also includes course materials provided by the teacher. Both are divided into fifteen units arranged principally by phoneme class. The corpus contains a total of about 5000 waveform files. The speech waveforms in the corpus have been labeled using the Toolkit forced alignment package described above, with human correction as needed. The label files use the Worldbet symbol set [9]. The concepts and corpus entries for each unit are summarized below:

1. *Introduction to sound*: Students study characteristics of common sounds. The corpus contains sine wave and tuning fork waveforms, music, animal noises, and mechanical and electronic sounds of daily life.
2. *Speech*: Students learn the IPA, Worldbet, and elementary labeling, plus the five phonation types: voicing, plosion, friction, aspiration, and silence/noise.
3. *Introduction to vowels*: Students learn about the acoustic tube model of vowel production and the three quantal vowels (Worldbet /i:/, /A/, and /u/). The corpus includes these vowels pronounced by a range

- of speakers at different loudness, intonation, and pitch levels.
4. *Introduction to consonants*: Student learn about four major consonantal phonation types: aspiration is represented by /h/, frication by /s/, plosion by /th/, and nasal phonemes by /n/. Coarticulation between the quantal vowels and these four consonants is explored. The first consonant clusters, /s t/ and /t s/, are encountered.
 5. *Monophthongs*: Students learn about the vowel trapezoid with its front/central/back and high/mid/low axes, and about coarticulation effects between the vowels and the four consonants above.
 6. *Plosives*: Students learn to distinguish the closure, release, and aspiration phases of plosives. The continuum between fully voiced and voiceless obstruents is sampled. The voicebar is first encountered during plosive closure. The bilabial and velar places of articulation are introduced. Plosive coarticulation with vowels, and anticipatory and perseverative patterns of vowel coarticulation in the presence of plosives are introduced by place of articulation. Initial, medial, and terminal plosives are contrasted. Coarticulation with /s/ is further explored.
 7. *Diphthongs*: Students learn about the six moving vowels of American English, divided into two categories according to whether the tongue movement is toward the front of the mouth or the back. Concepts include the characteristics of moving formants, the major and minor nuclei, and the atypical diphthongs /ei/ and /iU/.
 8. *Fricatives*: Students compare the voiceless alveolar fricative /s/ with the voiced /z/, extending their notions of voicing-obstruent interaction. The four American English fricative places and gestures of articulation are contrasted: the already familiar alveolar and the newly introduced labiovelar, interdental, and palatoalveolar.
 9. *Nasals*: Students study the separation of formants into oral and nasal subformants, the appearance of the nasal formant in vowels beyond the nasal itself, and the similarities between nasals and the corresponding homorganic plosives.
 10. *Approximants*: Students learn about the glides and liquids, light and dark /r/ and /l/, devoicing of approximants after obstruents, and the numerous English consonant clusters involving the approximants.
 11. *Other phonemes and allophones*: Students study the affricates /tS/ and /dZ/, the flap as a plosive allophone, voiced /h/, syllabic /n/ and /l/, and other important allophonic forms.
 12. *Syllables*: Students learn to recognize major prefix, root, and suffix morphemes in American English as a unit rather than as a sequence of constituent phonemes.
 13. *Words*: Students have access to a large corpus of words. Crucial is the ability to request a random word from the corpus and identify the word without listening to it. The word corpus includes more than four thousand American English words pronounced by a variety of male and female speakers, some with foreign accents.
 14. *Phrases*: Students study the rudiments of normal connected speech, including stress patterns, the different treatment of content versus function words, and examples of common phonological rules. The corpus includes two hundred short phrases pronounced by one male and one female speaker.
 15. *Introduction to other corpora*: Students learn about and view data from several major online corpora. The spectrogram reading corpus includes fifty example utterances drawn from various OGI corpora including numbers, letters, foreign language, telephone, and cellular speech.
- The order of the phoneme-oriented units was carefully chosen to present first those phonemes and classes easiest to recognize and distinguish from one another, and to end with those phoneme classes and combinations which are most difficult to read. For example, the weak fricatives are introduced late in the course because of their low energy, while approximants are introduced last among important phoneme classes because of their great plasticity.

4. Future course development

The speech tools and corpus described in the previous two sections require a human instructor. In order to make these materials available to a wider audience, we are directing our current work toward a completely computer-based spectrogram reading course, available as a future component of the CSLU Toolkit. Progress toward this goal has involved the development of three new Toolkit modules as described below.

4.1 Three-dimensional spectrograms

SpeechView has been enhanced by the addition of OpenGL 3D spectrogram displays as an alternative to traditional 2D gray-scale images. This module was undertaken specifically to enhance the readability of spectrograms for the course, and is now complete. In 3D mode, each second of speech is modeled by up to 100,000 polygons forming a Gouraud-shaded surface in 3-space. Both black and white and color 3D spectrograms use a combination of saturation and apparent surface height to achieve redundant visual coding of spectrogram energy; color 3D spectrograms additionally use color hue to code for the most prominent spectral peaks at each moment in time. These 3D spectrograms were used in the most recent spectrogram reading class, and instructor and class alike found them superior to traditional gray-level spectrograms.

4.2 Improved labeling capabilities

The second new Toolkit module involves a new speech labeling standard. Existing label files at OGI, and indeed all label files known to the authors, describe one-dimensional segments along the time axis. With this type of labeling, it is up to the instructor to describe the patterns that are contained within each labeled interval, by verbal description and by pointing to relevant shapes in the spectrogram. It is clear that 2D spectrogram labeling offers more precise labeling capabilities than traditional 1D labeling. Patterns of interest span variously shaped regions in the time-frequency plane, and 2D labeling can capture these region boundaries more accurately than 1D labeling.

To our knowledge, no 2D labeling standard currently exists. We have created the first draft of such a standard, including a set of labels based on the Worldbet symbols, but with extensions for phonation regions and the 2D patterns which characterize speech at various scales. This draft standard is available online [2]. We are interested in feedback from all interested parties.

4.3 Course authoring tools

The third Toolkit component required for any online course is a set of authoring tools to sequence and display course text and graphics, and to link these course materials to relevant waveforms, spectrograms, and both 1D and 2D label displays. We are in the process of developing such authoring tools for general use in building courses with the CSLU Toolkit. These authoring tools include an HTML-compatible tag set, hyperlinks between the various course units and resources, glossary entries, questions and other interactive exercises, and dynamic speech displays illustrating the course concepts. These tools and course materials are not yet complete, but a prototype course is on display at this conference and via the Internet [2].

We anticipate availability of the complete course in a Toolkit release in the summer of 1999.

5. Acknowledgements

This article is dedicated to Victor Zue, who demonstrated the feasibility of decoding speech from spectrographic displays and shared this knowledge with a large community of researchers, educators, and practitioners through publications and spectrogram reading courses. Through his work, the speech research community understands the importance of using acoustic phonetic knowledge to design spoken language systems. The spectrogram reading course described in this article is a direct result of his work.

SpeechView was developed by John-Paul Hosom. It builds on the efforts of many people, including Fil Alleva, Mark Fanty, Pieter Vermeulen, and Tim Carmell.

This work was supported in part by ONR/DARPA grant N00014-94-1-1154, NSF CARE grant EIA-9996075, and NSF CHALLENGE grant CDA-9726363. The views

expressed in this paper do not necessarily represent the views of ONR or NSF.

6. References

- [1] Zue, V (1985). *Notes on Spectrogram Reading*, Dept of EECS Technical Document, MIT, Cambridge.
- [2] <http://cslu.cse.ogi.edu/toolkit/specread>
- [3] Sutton, S., Cole, R., et al. (1998) Universal Speech Tools: the CSLU Toolkit, *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 3221-3224, Sydney, Australia, November 1998.
- [4] Massaro, D. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press, Cambridge, 1998.
- [5] Black, A., Taylor, P. (1997) *Festival Speech Synthesis System: System documentation (1.1.1)*, Human Communication Research Centre Technical Report HCRC/TR-83, Edinburgh.
- [6] Ladefoged, P. (1993). *A Course in Phonetics*, 3rd Edition, Harcourt Brace College Publishers, Fort Worth, 1993.
- [7] Cole, R.A., Rudnicky, A. I., Zue, V.W., and Reddy, R., (1980) Speech as Patterns on Paper, In R.A. Cole (ed.) *Perception and Production of Fluent Speech*. Hillsdale, N.J., Lawrence Erlbaum Associates, 1980.
- [8] *Speech Production and Perception I*, Software from Sensimetrics Corporation, Cambridge, 1997.
- [9] Hieronymus, J. (1994). *ASCII phonetic symbols for the world's languages: Worldbet*. AT&T Bell Laboratories, Technical Memo.