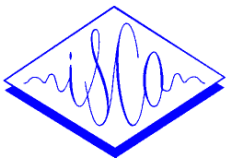

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

CSRE: A SPEECH RESEARCH ENVIRONMENT

Donald G. Jamieson[1], Ketan Ramji[1], Issam Kheirallah[1], and Terrance M. Nearey[2]

[1]Speech Communication Laboratory, Hearing Health Care Research Unit
Department of Communicative Disorders, The University of Western Ontario
London, Ontario
and
[2]Department of Linguistics, University of Alberta
Edmonton, Alberta

10.21437/ICSLP.1992-333

I. ABSTRACT

CSRE (The Canadian Speech Research Environment) is a comprehensive, microcomputer-based system designed to support speech research using IBM/AT-compatible microcomputers [4]. CSRE provides a powerful, low-cost facility in support of speech research, using mass-produced and widely-available hardware. The project is non-profit, and relies on the cooperation of researchers at a number of institutions. Work on the project is supported primarily by the fees generated when the software is distributed. Version 3.0 of CSRE has been used since 1989 by researchers in more than 100 laboratories in 12 countries. Version 4.0 offers a wider range of functions, runs faster, uses higher resolution displays, and supports additional hardware systems, including digital signal processing boards. Functions include speech capture, editing, and replay; several alternative spectral analysis procedures, with color and surface/3D displays; parameter extraction/tracking and tools to automate measurement and support data logging; alternative pitch-extraction systems; parametric speech (KLATT80) [7] and non-speech acoustic synthesis, with a variety of supporting productivity tools; and a comprehensive experiment generator, to support behavioral testing using a variety of common testing protocols.

II. INTRODUCTION

CSRE (The Canadian Speech Research Environment) provides a comprehensive, integrated, inexpensive, microcomputer-based system to permit speech researchers: 1. to record, store, and play back high quality natural speech signals; 2. to edit stored natural speech signals -- to cut and otherwise modify portions of the signal and to concatenate (paste together) different signals; 3. to analyze and accurately measure the frequency, amplitude, and duration of speech and other acoustic signals; 4. to parametrically synthesize speech and other complex acoustic signals; and 5. to control speech output for listening tests or experiments.

Earlier releases of the CSRE system implemented the basic functions which speech researchers require, using mass-produced, widely-available hardware [4]. Release 4.0 has optimized these functions, added new functions, improved our productivity tools, and improved the interactive environment and the speed with which functions are executed.

The system requires an IBM/AT compatible computer (80386 or higher system strongly recommended), with at least VGA graphics, mouse and audio input/output system consisting of a board with a digital-to-analog converter (eg., Ariel's DSP-56, or DSP-16, Tucker-Davis Technologies Turnkey system (available Spring 1993), or Data Translation DT2801A or DT2821), microphone, preamplifier and filters. The software is available to all interested researchers.

III. COMPONENT PROGRAMS OF THE CSRE PACKAGE

III.A. Waveform Sampling and Editing

The Waveform Editor is used to digitize, edit and manipulate speech signals within the time-domain. The signal of interest may be taken from a disk file, or recorded directly from microphone or tape input.

Mouse-driven menu commands are used to select functions -- for example to display a signal, to isolate a segment of the displayed signal, to play a signal or segment, to increase or decrease the amplitude of a signal or segment, or to cut the selected signal and save it to a disk file. At any time, the signal in any window may be played, or a window may be positioned on any portion of the signal and the windowed portion played.

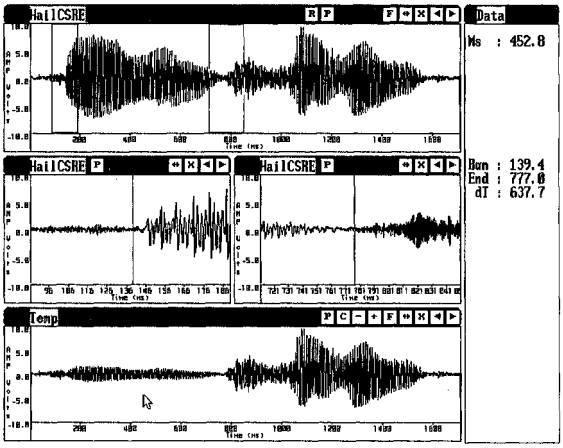

Figure 1. Sample Waveform Editor screen. A sampled waveform (amplitude by time) is displayed in the top portion of the screen, with the results of the same waveform, after editing, displayed in the bottom portion of the screen. The mouse is used in the top portion of the screen to define the region to be edited, in terms of two windows. The left window contains the start of the edit segment, while the right window contains the termination of the edit segment. These two windows are displayed in the middle screen, to the left and right, respectively. A vertical cursor is set by mouse control in the left middle window to define the exact beginning point for the edit segment; a similar cursor is set in the right middle window to define the end of the edit segment. In this example, the editor has been used to digitally attenuate the windowed segment, reducing its amplitude relative to the remainder of the signal. Other options include cutting and digitally amplifying the segment.

III.B. Speech Analysis

III.B.1. Analyses Procedures

Spectral analysis procedures estimate the true time-frequency-amplitude characteristics of a complex signal. In CSRE 4.0, we have provided four spectral estimation techniques, offering researchers a range of analysis options. It is important to do this because all available procedures have some advantages and some disadvantages. The procedures available within CSRE 4.0 are the fast Fourier transform (FFT), two autoregressive techniques -- the autocorrelation (AC) and modified covariance (MC) methods -- and the Cone-Kernel (CK) approach.

FFT spectral estimates are characterized by many tradeoffs in windowing, time-domain averaging, and frequency-domain averaging. These balance the need to reduce sidelobes, to perform effective ensemble averaging, and to ensure adequate spectral resolution. This technique is computationally efficient, but has many problems associated with the assumptions regarding data outside the measurement interval. In addition, frequency resolution is limited to the reciprocal of the time domain data length [5].

Autoregressive methods have been designed as alternative spectral estimation procedures that overcome some of the inherent limitations of the FFT approach. Rather

than assuming that the data outside the window are zero, the spectral estimates are based on the power spectrum density implied by a model which approximates the actual underlying process. The need for windowing functions can be eliminated, along with their distorting impact. As a result, the improvement over the conventional FFT spectral estimate can be quite dramatic, especially for short records [5].

Speech signals are well approximated by autoregressive models because of the close relation of this modelling approach with the linear prediction analysis, where a speech sample is approximated as a linear combination of previous speech samples. In the autocorrelation method, the signal is windowed and then padded with zeros before the autocorrelation coefficients are calculated. AC is successful, in part, because it generates an all-pole model which is guaranteed to be stable. However, the approach provides limited spectral resolution, as closely spaced spectral lines cannot be resolved with a short data window. Furthermore, the windowing decreases resolution and introduces spectral distortion. In certain situations, the AC method also produces spectral line splitting, where two or more peaks occur when only one peak should be present in the spectral estimate [6]. Finally, the AC approach models the peaks in the spectrum better than the valleys.

In the *Modified Covariance* method, the operations are carried out on the data directly without any zero padding. The MC method utilizes a combination of forward and backward linear prediction for estimation. The MC method works better for short data segments, offering sharper response for narrowband processes. While the covariance technique operates on the data directly, the AC approach uses the biased autocorrelation estimates to reduce the risk of ill-conditioning, but at the expense of a degradation of the autoregressive spectral resolution and a shifting of spectral peaks from their true locations [10]. In addition, no spectral line splitting is observed with the modified covariance method. This approach can exactly match an all-pole spectrum from a short section of data, but stability cannot be guaranteed.

While the three techniques discussed above represent current conventions for speech analysis and work well in many analysis situations, they have at least three limitations: 1) they are quasi-stationary approaches, being applied to analyze speech and other nonstationary signals; 2) the FFT permits good time resolution or good frequency resolution, but not both simultaneously; and 3) the autoregressive techniques use an all-pole model to represent the signal, but this model is known to be incorrect for unvoiced segments of speech, and for speech embedded in noise.

The bilinear time-frequency distributions offer the possibility of analyzing nonstationary signals. The first of these was the Wigner distribution [eg., 1,2,3]. However, when applied to speech and other multicomponent signals, the Wigner distribution becomes extremely difficult to interpret because of the interfering cross-terms which result from the nonlinear nature of this distribution. Post-processing can reduce these interfering terms without sacrificing the desirable properties of the Wigner distribution. In particular, the *Cone-Kernel* (CK) method minimizes the artifact problem of the Wigner distribution, permitting spectrograms of speech signals, based on high-resolution analyses in both time and frequency. The basic principle behind the CK approach is that the time supports of the kernel in each direction of a two-dimensional time plane are made independent. The kernel takes the form of a cone. The CK time-frequency representation is obtained by convolving the cone-shaped kernel with the signal correlation and taking a Fourier transform. The resulting representation simultaneously preserves the property of finite time support, enhances spectral peaks and smooths the cross-terms [11]. The CK technique is suitable for analyzing speech signals where there is a need to resolve two closely-spaced formants, or for tracking a rapidly-changing spectral peak. Unlike a narrowband spectrogram, however, this improved frequency resolution is obtained *without* sacrificing time resolution. This makes it possible to obtain a good estimate of the time of occurrence of an event. Thus, the CK approach provides information about speech which previous techniques have not been able to offer. An example is the case of spectral peak splitting observed in the first formant over a single pitch period for voiced speech [8].

III.B.2. Spectrum Displays and Interaction Tools

Within CSRE, the primary display of the results of a spectrographic analysis is in the form of a perceptually-compelling color spectrogram, with time on the horizontal axis, frequency on the vertical axis and amplitude as one of 256 possible colors to code the amplitude of the signal. To facilitate analyses, the user can always play back the full signal displayed in the spectrogram or any desired portion of the spectrogram display. Numerical values of time, frequency and amplitude and a full spectral slice associated with any point on the spectrogram can be displayed using the mouse. The frequencies and amplitudes of the major peaks for this slice are automatically extracted; any value in the slice can be read from the display and logged to disk.

A wide range of control over the spectrogram display is available. The signal can be scrolled through the spectrogram window, to study earlier or later parts of the signal. The displayed signal can be zoomed, to increase or decrease the range of times and/or frequencies displayed. The user can switch to other displays, containing other analyses (the time-domain signal, pitch, or amplitude). The dB range represented by each color is shown as part of the display, and the range of dB values displayed can be altered by the user. Several alternative mappings from intensity to color can be selected: a rainbow scale, a temperature (heat) scale, and a brightness (gray) scale. Inexpensive high-quality color hardcopy output is straightforward, using an HP PaintJet printer, with a print utility such as Pizzaz Plus. The same print utility supports high-quality 16/64 grayscale output, suitable for publication.

For any desired sampled signal, the power spectrum density is calculated and a spectrogram can be generated under menu control. Each parameter has a default value, but users can override this default to specify: 1. the desired frequency resolution (number of filters); 2. the desired temporal resolution; 3. the lower and upper frequency boundaries of the spectrogram (in Hz); 4. the type of analysis procedure to be used; 5. the color scale to be used; 6. the range in dB (subset of colors) to be displayed; 7. whether the intensity of the spectrogram is to be scaled *relative* to the within-signal maximum value, or *independently* of signal intensity -- with scaling relative to the maximum signal which could be recorded using a 12 (or 16) bit analog-to-digital converter; and 8. whether or not the signal is to be preemphasized.

Once a spectrogram has been displayed, a mouse-controlled cursor can be used to define and play any desired region of the signal, to obtain a direct readout of the time, frequency, and amplitude coordinates of any desired point, or to select and display a "spectral slice" (amplitude-by-frequency display) at a particular time point in the signal. A mouse- or key-controlled cursor can also be used to obtain amplitude values for any desired frequency in the slice.

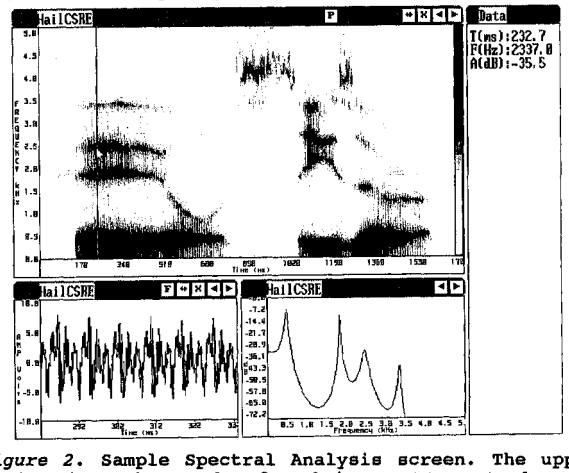

Figure 2. Sample Spectral Analysis screen. The upper window shows the result of applying an AC spectral analysis to the utterance "hail Caesar". Time is represented on the horizontal axis, frequency on the vertical axis, and amplitude as the darkness of the display. Normally, amplitude is displayed in CSRE as one of 256 possible colors on the CRT screen or in hard copy as 64 gray-levels or colors; here, the photoreduction technique reduces amplitude resolution to approximately 8 gray-levels. Any desired portion of the signal displayed in the spectrogram can be played back under mouse control. The mouse can also be used to move a pointer to obtain measured values from the signal; in the example,

the pointer identifies a point 232 ms from the start of the signal, where the frequency is 2337 Hz and the amplitude is -35.5 dB, relative to the loudest part of the signal. The first four formants of the signal can be seen in the associated spectral slice centered at 232 ms (see bottom right window). The bottom left window displays the time waveform for a brief windowed period centered at the same point.

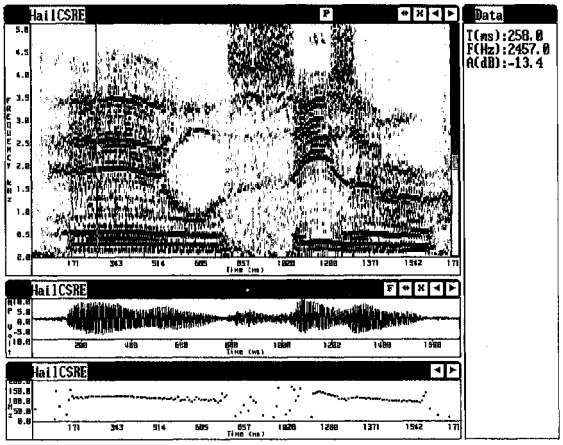

Figure 3. Sample Spectral Analysis screen. The upper window shows the result of applying a Cone-Kernel spectral analysis to the utterance "hail Caesar". The signal, dimensional representation, playback control and measurement options are as described in Figure 2 for the AC method. However, the Cone-Kernel procedure reveals new acoustic details, including the formant transitions away from the dominant vowels (compare Figure 2). The middle and bottom windows display the time waveform and a pitch track (obtained using the comb-filtering approach) for this utterance, respectively, using the same time scale.

III.B.3. Formant Tracking

Basic formant tracking in CSRE uses linear predictive analysis procedures similar to those described by Markel and Gray (1976). The time-varying signal is approximated by a series of overlapping, short-term analyses. At each step, the signal is multiplied by a Hamming window and pre-emphasized to ensure that there is sufficient energy in the higher formants. Autocorrelation coefficients are then estimated from a selected portion of the resulting signal. The Levinson-Durbin algorithm [9] is then used to solve for the LPC coefficients, given the autocorrelation coefficients. A peak-picking procedure is used to extract peaks of the power spectrum density of the autoregressive model used. On output, formant candidates are identified in frequency-sorted order. Candidate formant tracks are displayed, together with estimates of the bandwidth and amplitude of each formant at each time point.

III.B.4. Pitch Extraction

Two pitch extraction algorithms are provided with CSRE 4.0: a modified Cepstrum-based procedure and a Comb-Filtering method. In the Cepstrum procedure, the signal is first multiplied by a Hamming window. A log magnitude spectrum of the Cepstral segment is then calculated. The spectrum is windowed by a three-piece function: an increasing half cosine taper from zero to 200 Hz, the identity function from 200 to 1000 Hz and a decreasing half cosine taper from 1000 to 1500 Hz. This step improves the results when noise is present or the speech is distorted. The Cepstrum is computed on this adjusted, windowed spectrum by taking a second FFT. The Cepstral coefficients are weighted by a function which is zero outside the Cepstral frequency index ranges. This weighting function decreases the dependency between the Cepstral peak value and the fundamental frequency. The Cepstral bin with the highest weighted value is selected and the peak of the weighted Cepstrum is found. The fundamental period is estimated by quadratic interpolation of the three points surrounding the pitch peak and is converted to a pitch estimate.

In the Comb Filtering pitch estimation procedure, the speech signal is first divided into blocks. The data within each block are then low-pass filtered with an FIR filter. The frequency of the first and second formants are then estimated, and F1 and F2 are eliminated by Comb Filtering. The resulting signal is smoothed by sequential median averaging, followed by short-term arithmetic averaging. Pitch extraction is applied on the resulting signal, using two criteria: threshold-passing followed by positive-to-negative zero-passing.

All pitch estimation procedures have limitations, and will fail in some circumstances. Our Comb-filtering procedure for pitch extraction offers the user the option of examining the individual pitch estimates on a cycle-by-cycle basis to accept or reject the individual estimates provided by the system. This process involves selecting a portion of the pitch track to be displayed in detail. The "detail" display contains each individual estimate of the glottal cycle, superimposed on the smoothed waveform. The original (unprocessed) time-domain waveform is also available for viewing. Users may edit the markers on the smoothed waveform, by adding a marker at a given point on the waveform, deleting a marker judged to have been placed extraneously, or moving a marker from one point to another in the waveform. In our experience, this two-step process allows meaningful pitch tracks to be obtained with confidence for acoustic speech waveforms where automatic pitch extraction alone would fail.

III.B.5. Amplitude Extraction

An amplitude extraction procedure consisting of computing the RMS value of the data contained in a sliding window is used to provide an amplitude track. The results of this analysis can be displayed on the same time axis as other analyses (eg., formant track, pitch track, time waveform, or spectrogram), as desired.

III.C. Synthesis

III.C.1. Parametric Speech Synthesizer

CSRE provides a parametric speech synthesizer which generates *audio data* files from the specifications of 40 control parameters -- the 39 parameters of the KLATT80 synthesis [7], plus parameter 40, CORRSW, added by Kewley-Port. In this implementation, the synthesizer is controlled by an INTERPolated file, which specifies the values of each parameter, updated every 5 ms. The INTERPolated files are generated from KNOT files, which define "skeletonized" parameter tracks, specifying "piecewise linear" functions of the desired parameter tracks.

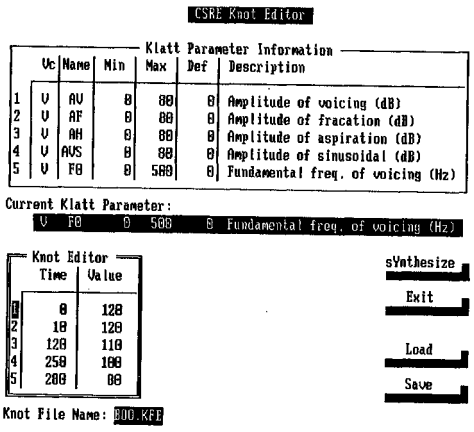

Figure 4. Sample screen showing input used to specify parameters to control the KLATT80 parametric speech synthesizer. Each of 39 control parameters for the synthesizer is defined at each desired change point (KNOT) within the signal. These KNOTS specify the desired parameter tracks as piecewise-linear functions; these KNOT files are fed to an interpolator to generate the complete files, prior to synthesis.

III.C.2. Continuum Generator

Many studies using synthesized speech involve, at one stage or another, an exploration of the perceptual effects of systematically varying the synthesis parameters. Typically, stimuli which bear such a parametric relationship to each other are conceptualized in terms of a "continuum" in which adjacent stimuli differ by a step in one (or more) parameter(s). Using conventional procedures, the creation of such a parametrically-varied continuum is slow and requires a high degree of attention, if errors are to be avoided; moreover the task is distinctly uninspiring. To facilitate such explorations, we developed a "Continuum Generator", in the form of a

restricted, high-level, programming language.

In its simplest application, the Continuum Generator allows a researcher to specify a range of values to be assumed by a variable -- the starting value, number of steps, and size of the steps -- and then to synthesize all of the speech signals which would result from the combination of each of the possible values of the variable parameter, in turn, together with a specified set of fixed values for all of the other parameters which are required to control the synthesizer. In this way, the Continuum Generator operates much like a loop in a conventional programming language such as Basic. By extrapolation, nested loops can also be generated, resulting in the full set of synthesized speech signals produced by the factorial combination of two (or more) sets of variable parameters.

Using these procedures, we have been able to quickly generate large numbers of speech signals for use in speech perception experiments, with a minimum of effort, and with the details of the synthesis parameters controlled automatically (and without error). Indeed, some experiments using these procedures involve several hundred parametrically-synthesized signals -- offering new opportunities to explore combinations of parameters which would previously have been impossible because of the time constraints of traditional synthesis techniques. One application in which the Continuum Generator has proved to be of particular value involves the study of interactions between parameters; another valuable application involves the identification of combinations of parameters which result in very high quality synthesized speech.

III.C.3. Nonspeech Acoustic Signal Synthesizer

The Acoustic Signal Synthesizer is a general-purpose, interactive program which permits users to generate steady-state or swept acoustic waveforms, defined in terms of a variety of parameters including source (cosine, square, or triangle waveform), envelope, signal duration, sampling frequency, and initial and final signal frequencies. The interface is menu-based; facilities include the ability to concatenate two or more signals which have been generated with the system. The resulting waveforms can be used with other parts of the CSRE system.

III.D. Experiment Generator

The Experiment Generator (EGEN) permits one to quickly define and test listeners in certain familiar types of identification and discrimination experiments. Audio data files generated by CSRE, either through sampling and editing of real speech, or through parametric synthesis, are presented to listeners in a specified sequence and according to a specified timeline. The user's responses are recorded automatically by the computer, and logged to disk file for later analysis.

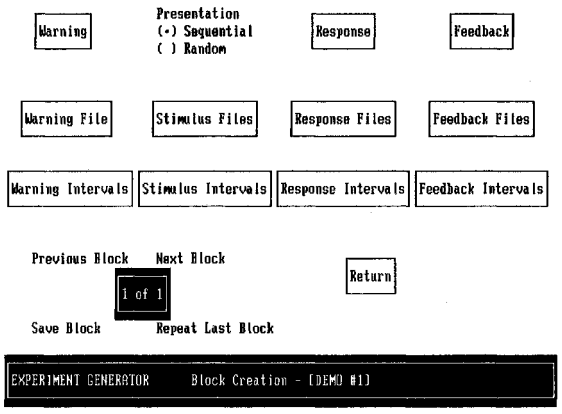

Figure 5. Sample Experiment Generator screen used to define blocks of a listening test. By selecting the indicated screen "button" with the mouse, the user can define the critical aspects of the listening test, including whether or not a warning signal is to be played and when, which stimuli are to be included ("Stimulus Files"), and with what interstimulus delays ("Stimulus Intervals"), what responses are to be offered to the listener as either screen text or graphics ("Response Files") and over what intervals these are to be available, whether and what kind of feedback is to be given after the listener responds.

EGEN is menu-driven and permits an audio experiment to be defined by filling in a series of forms specifying such details as the names of the audio stimulus files to be used in the experiment, the number of times each signal is to be presented in each block, the sequence of events within each trial and the rate of presentation, the time permitted for the subject to respond, the intertrial interval, the number of blocks to present in a session, whether stimuli are to be presented in a randomized order within a block or presented in a predetermined sequence, and so forth. Once these details have been specified, the program will produce an experiment specification file which can be read and used by our Experiment Controller to carry out the experiment specified. This approach makes it possible to define and carry out the most common types of listening experiments, using different stimulus sets, quickly and easily.

ACKNOWLEDGEMENT AND AVAILABILITY

The contributions of Cathy Mandarino, Peter Bangarth, Emmet Raftery, Todd Schneider, Terry Baxter, Tom Welz, and Pierre Divenyi, and the support of the Natural Sciences and Engineering Research Council of Canada and Bell Northern Research in helping to launch the CSRE project, are gratefully acknowledged. The CSRE project is non-profit and cooperative, with development and distribution costs covered by the funds provided when researchers purchase CSRE. The current (1992) price for the CSRE 4.0 system is US$500 for new users; CSRE 3.0 owners may upgrade to CSRE 4.0 for US$200. Those interested in obtaining the system should contact Dr. D.G. Jamieson at the indicated address, or by e-mail (JAMIESON@UWOVAX.UWO.CA).

REFERENCES

[1] T.A. Claasen & W.F. Mecklenbrauker. "The Wigner distribution - A tool for time-frequency signal analysis. Part 1". *Philips Journal of Research, 35*, 217-50, 1980.

[2] T.A. Claasen & W.F. Mecklenbrauker. "The Wigner distribution - A tool for time-frequency signal analysis. Part 2". *Philips Journal of Research, 35*, 276-300, 1980.

[3] T.A. Claasen & W.F. Mecklenbrauker. "The Wigner distribution - A tool for time-frequency signal analysis. Part 3". *Philips Journal of Research, 35*, 372-89, 1980.

[4] D.G. Jamieson, K. Ramji & T.M. Nearey. "CSRE: A Speech Research Environment". *Canadian Acoustics, 17*, 23-35, 1989.

[5] S.M. Kay & S.L. Marple Jr. "Spectrum analysis - A modern perspective". *Proceedings of the IEEE, 69*, 1380-419, 1981.

[6] S.M. Kay & S.L. Marple Jr. "Sources of and remedies for spectral line splitting in autoregressive spectrum analysis". *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, 151-4, 1979.

[7] D.H. Klatt. "Software for a cascade/parallel formant synthesizer". *Journal of the Acoustical Society of America, 67*(3), 971-95, 1980.

[8] P.J. Loughlin, L.E. Atlas, & J.W. Pitton. "Advanced time-frequency representations for speech processing". In M. Cooke & S. Beet (eds) *Visual representations of speech analysis*, Chichester: J. Wiley, 1992.

[9] J.D. Markel & A.H. Gray. *Linear prediction of speech*. New York: Springer-Verlag, 1976.

[10] S.L. Marple Jr. "A new autoregressive spectrum analysis algorithm". *Transactions on Acoustics Speech and Signal Processing, 28*, 441-51, 1980.

[11] Y. Zhao, L. Atlas & R. Marks II. "The use of cone-shaped kernels for generalized time-frequency representations of nonstationary signals", *IEEE Transactions on Acoustics Speech and Signal Processing, 38*, 1084-91, 1990.