An On-Line Laboratory Course on Speech Analysis

VAGNER L. LATSCH,[1] FERNANDO G. V. RESENDE, JR.,[1] SERGIO L. NETTO[2]

[1]DEL/EE, *Federal University of Rio de Janeiro, Rio de Janeiro*, RJ, *Brazil*

[2]*Program of Electrical Engineering*, COPPE, *Federal University of Rio de Janeiro, Caixa Postal* 68504, *Rio de Janeiro*, RJ, 21945-970, *Brazil*

Received 5 May 2000

ABSTRACT: An on-line laboratory course for speech signal processing is described. The course consists of a series of practical experiments involving the three aspects of speech processing: coding, synthesis, and recognition. The experiments were developed using an auxiliary tool implemented in Delphi and were then transformed into individual class-modules using the eTEAM software. Both pieces of software are briefly described, along with three experiments already available through the Internet. © 2000 John Wiley & Sons, Inc. Comput Appl Eng Educ 8: 178–184, 2000

Keywords: speech signal processing; on-line course; computer laboratory experiments

INTRODUCTION

Speech signal processing has become in the last decades a very active research area. Such development came about greatly due to two main aspects: first, the availability of new digital signal processor devices that are able to process data at extremely large speeds and affordable prices; secondly, the surge of a great consumer market thirsty to purchase all sorts of electronics devices in today's new technology age.

Teaching a speech processing course usually follows a few courses on "signals and systems," "digital signal processing," and sometimes "probability and random signals." Also, standard courses and textbooks often divide the subject into three main topics, namely, speech coding, speech synthesis, and speech recognition.

Speech coding is the process that transforms the speech signal in a format suitable for transmission or recording, often occupying much less bandwidth or disk space than the original signal. We find such technology in telecommunication systems, including mobile phones and digital computers, for example.

Speech synthesis is the process of generating audible speech with a digital machine, usually starting from written text. Such technology has become very useful to seeing-impaired people, for instance, in many ways.

Finally, speech recognition consists of making a machine understand human speech and react to that in a specific form. Applications of such technology include, for example, speech-to-text automatic conversion and voice-operated machines as toys and mobile phones.

Although the applications of speech processing include a wide variety of practical systems, teaching such techniques can be rather dull. In fact, a real

Correspondence to S. L. Netto (sergioln@lps.ufrj.br)

Contract grant sponsor: FUJB-UFRJ; contract grant number: 8450-6.

Contract grant sponsor: CAPES.

Contract grant sponsor: CNPq; contract grant number: 520187/98-9.

understanding of speech processing is only achieved through experimentation, as optimal systems are developed mainly with extensive research. One must listen to a speech-based system to truly understand its capabilities, and fully perceive why it performs the way it does [1].

For that matter, a good course on speech signal processing must include a strong theoretical background accompanied by a series of computer experiments that greatly stimulate the students to learn. This paper thus presents the structure of a speech signal-processing course strongly based on computer experiments. The final version of the course is directed to distance-learning as its laboratory modules are developed with the help of two software tools: the Speech-Analysis Program (SAP), developed at the Department of Electronics Engineering at the Federal University of Rio de Janeiro; and the well-known eTEAM software, specific for distant learning over the Internet.

This paper is organized as follows: in Introduction to Speech Analysis, we describe some basics of speech signal processing that will be addressed by the proposed course. In The Speech Analysis Program, we briefly present the SAP software, which implements the speech processing techniques mentioned in Introduction to Speech Analysis. In The eTEAM Software, the eTEAM software is described with emphasis on its capabilities that made it a very useful tool for distance learning. Finally, in Practical Experiments, the contents and formats of three laboratory experiments already implemented for the proposed course are described. These class-modules are made available at the Internet and can be downloaded at the URL: http://www.lps.ufrj.br/projects/proj19.html, which is the main Internet address for this project.

INTRODUCTION TO SPEECH ANALYSIS

Speech Segmentation

Common speech signals are bandpass signals occupying the frequency range of 20 Hz to 8 kHz, with, however, most of the energy concentrated in the range 50–500 Hz [2]. This energy-uneven concentration can cause numerical problems in some later processing stage of the speech signal, as in the case of linear prediction analysis seen below. To reduce this effect, a speech signal should be pre-processed by a one-pole highpass filter which flattens up the overall signal spectrum.

After that stage the signal is ready for being segmented. To understand the importance of such an operation, consider 3 s of a speech signal sampled at 8 kHz, thus comprising a total of 24,000 samples. Processing such amount of data is highly prohibitive for most practical real-time applications. Furthermore, signal quasi-stationarity is necessary for practical system modeling (spectral analysis). We must then break down the original signal into smaller parts, called segments, by multiplying it by a window function of the type seen in Figure 1.

In that manner, we generate a segment starting from $n = m$, and lasting for N samples. Using the specific window function represented in Figure 1 (the so-called rectangular window), however, signal discontinuities close to the window edges are introduced in the time domain. These discontinuities generate large ripples (Gibbs oscillations) in the frequency domain. To reduce such problems other types of windows can be used, such as Hamming, von Hann (Hanning), Bartlett, Blackman, and triangular [3]. In speech processing [1, 2, 4], the most commonly used window is by far the Hamming window, defined as

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right), & n = 0, 1, \ldots, N-1 \\ 0, & \text{otherwise.} \end{cases}$$

Linear Prediction Analysis

Choosing the proper segment size is a very important issue on the segmentation step. In fact, if N is chosen too large, nonstationarity becomes an issue. On the other hand, an N too small implies a large number of segments. A reasonable value of N thus must be able to compromise these two aspects. In practice, segments of 5–30 ms are used. For an 8 kHz sampling rate, for instance, these values correspond to 40–240 speech samples, respectively.

Data reduction can then be achieved by considering the linear prediction (LP) model for a speech signal $s(n)$, as depicted in Figure 2. In this figure, $x(n)$ represents the excitation which is either a pulse train or a white noise, depending on the signal $s(n)$ being either voiced or unvoiced, respectively. Also, the

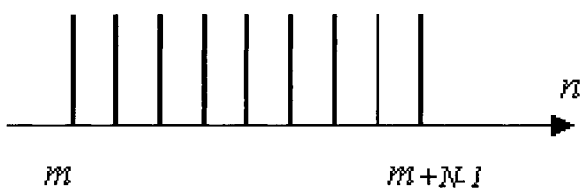

Figure 1 The rectangular window function of size N.

Figure 2 Linear prediction model.

autoregressive (AR) filter is a digital filter with a transfer function of the type

$$H(z) = \frac{G}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_M z^{-M}},$$

where G is the model gain, M is the model order, and the a_i's are the so-called LP coefficients. Practical values of M lie in the range $7 < M < 16$, and the computation of the LP coefficients follows from the normal equation [2]:

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_M \end{bmatrix} = \begin{bmatrix} r_s(0) & r_s(1) & r_s(2) & \cdots & r_s(M-1) \\ r_s(1) & r_s(0) & r_s(1) & \cdots & r_s(M-2) \\ r_s(2) & r_s(1) & r_s(0) & \cdots & r_s(M-3) \\ \vdots & \vdots & \vdots & & \vdots \\ r_s(M-1) & r_s(M-2) & r_s(M-3) & \cdots & r_s(0) \end{bmatrix}^{-1} \begin{bmatrix} r_s(1) \\ r_s(2) \\ r_s(3) \\ \vdots \\ r_s(M) \end{bmatrix},$$

where $r_s(k) = E[s(n)s(n-k)]$.

Cepstrum Analysis

The main advantage of the LP analysis is that it is able to separate the information concerning the vocal tract (which is represented by the AR filter in Figure 2) from the excitation signal. Another way of doing that is using the concept of cepstrum analysis. The cepstrum $c(n)$ of a signal $s(n)$ is defined as the inverse DFT of its magnitude spectrum in dB, or equivalently

$$c(n) = IDFT[\log|DFT[s(n)]|].$$

The main property of the cepstrum is that it isolates the pitch information (that characterizes the excitation of voiced signals) from the vocal tract information [2]. In that manner, we can model the vocal tract by retaining just the necessary information from the cepstrum $c(n)$. This is done by a process referred to as liftering, which corresponds to a lowpass filtering performed in the cepstrum domain.

THE SPEECH-ANALYSIS PROGRAM

The speech-analysis program (SAP) was developed at the Federal University of Rio de Janeiro, Brazil, as an educational tool for students of an undergraduate speech-processing course. The work was done in Delphi which is a high-level language for the Microsoft Windows. The main SAP v 1.2 interface screen is shown in Figure 3 below.

Some capabilities of the SAP in its present 1.2 version include:

- Processing. WAV (both 8- and 16-bit formats) and RAW files;
- Playing the speech sound before and after prefiltering, using any sound card connected to the Microsoft Windows 9× environments;
- Performing signal blocking with arbitrary segment size, and with or without overlapping between consecutive segments;
- Applying different kinds of window functions: rectangular (Boxcar), Hamming, von Hann, Bartlett, Blackman, and triangular, as can be verified in Figure 3;
- Performing LP analysis (for distinct values of the model order), extracting the LP coefficients, and plotting the resulting magnitude response;

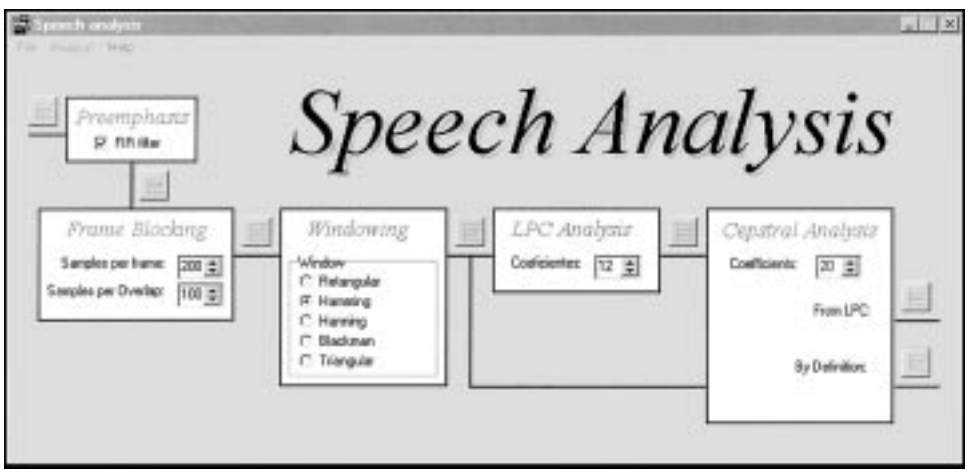

Figure 3 The SAP main interface screen.

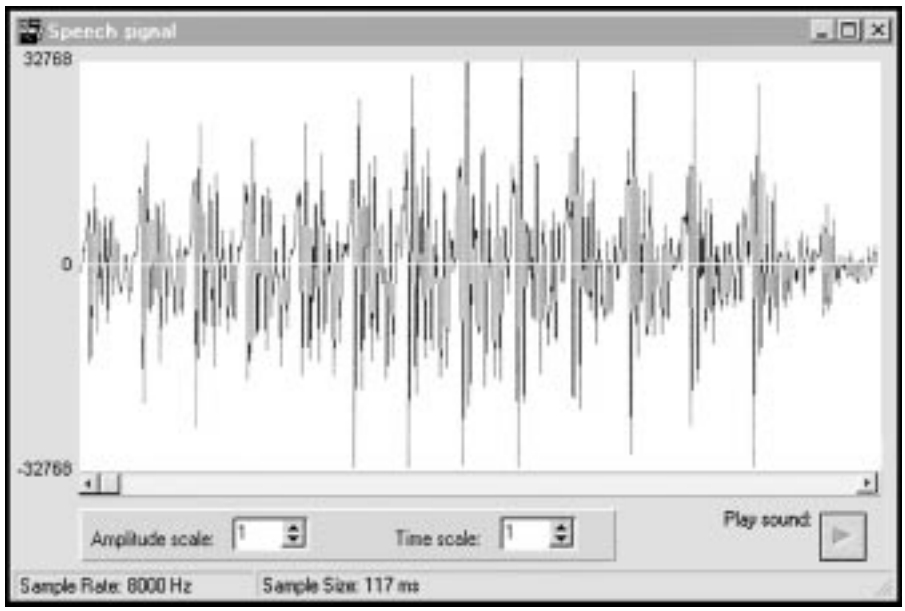

Figure 4 Speech signal visualization with the SAP.

- Performing cepstrum analysis, computing and plotting the cepstrum coefficients from the LP coefficients.

In Figures 4–7, we visualize some possible graphical outputs generated by the SAP software. Figure 4 depicts 40 ms of the sound of the vowel /a/ sampled at 8 kHz. Notice how easy it is to change the x and y axes for this kind of a plot. The button on the lower-right corner of the figure plays the respective sound in any sound card supported by the Microsoft Windows 9x operational systems.

Figure 5 shows a comparison of two 100-sample segments before (left-hand side) and after (right-hand side) performing the Hamming windowing on each segment. Notice how the segments on the right (after windowing) are smoother on their edges than the segments on the left.

Figure 6 depicts some results obtained with the LP analysis performed on a speech segment. The results include the AR model magnitude response and the corresponding coefficients for two LP analyses with $M = 5$ and $M = 30$. Notice that the model gain G is also provided by SAP in the two cases.

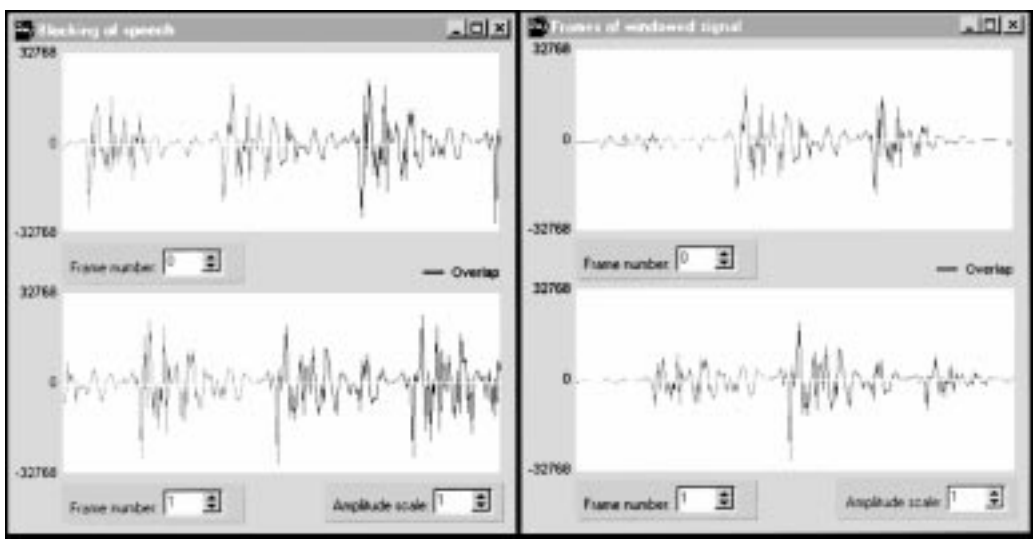

Figure 5 Visualizing speech segmentation and windowing with the SAP.

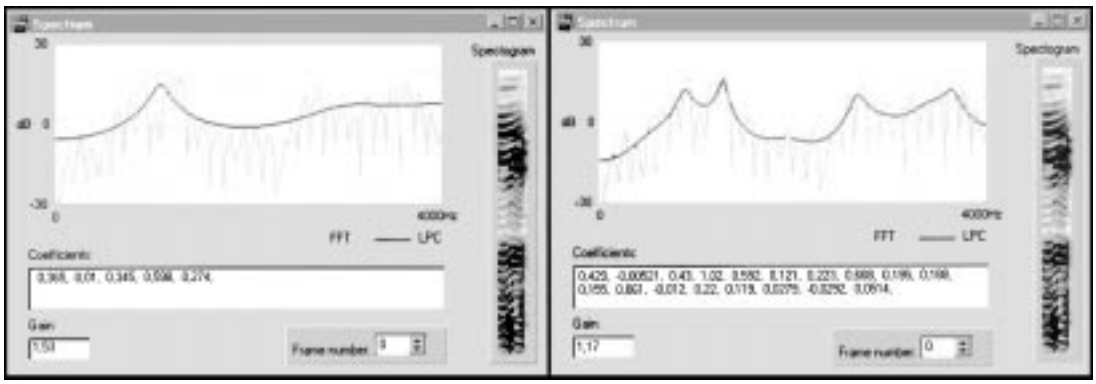

Figure 6 Magnitude response of LP models with $M = 5$ and $M = 18$ calculated with the SAP.

Finally, Figure 7 gives a sample on the possible results from the cepstrum analysis performed by the SAP. This figure shows the plain cepstrum coefficients $c(n)$ which correspond to a vocal-tract model distinct from the LP model. To obtain the magnitude response of the resulting cepstrum model, one must invert the operations described in the subsection Cepstrum Analysis.

THE eTEAM SOFTWARE

The eTEAM is a licensed software distributed by Infocast, Inc. The main eTEAM capabilities include the ability of managing different types of data in a single integrated environment. Such power makes the eTEAM extremely suitable for applications like distance-learning, as we are able to incorporate in each class module more information in several distinct formats, such as: figures, graphics, numbers, equations, audio, and video. All this allows a given speech-analysis course suited for the eTEAM software to be more dynamic, fully exploiting all aspects of speech signal processing. More information on the eTEAM package can be obtained at http://www.i-cast.net.

PRACTICAL EXPERIMENTS

This section presents three classes on speech analysis obtained by integrating the SAP and eTEAM tools.

Experiment 1: Signal Segmentation

The experiment starts by loading a given. WAV file on the SAP software. The whole signal is then shown to the student, as in Figure 4, so he/she can get familiarized with how a speech signal looks like. After that, through an audio explanation the student is told of the importance of breaking down the entire speech signal

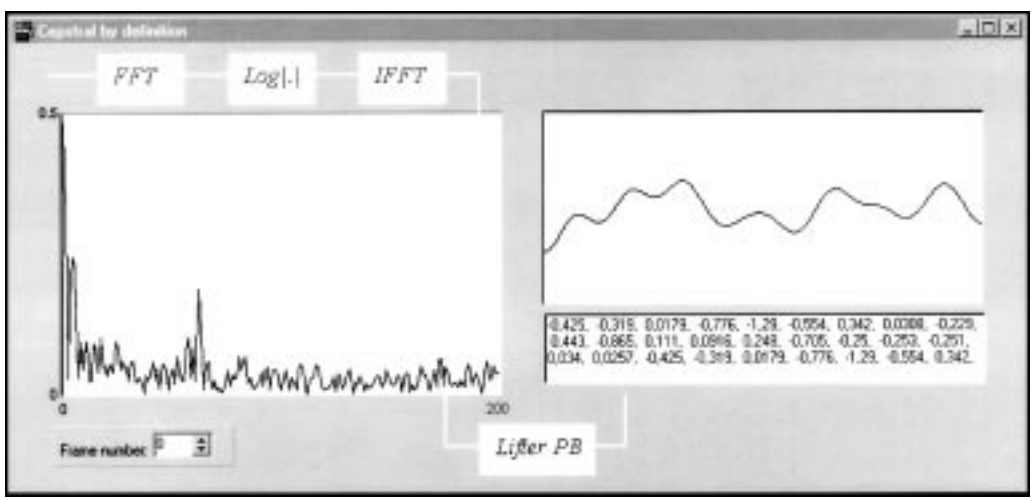

Figure 7 Cepstrum coefficients calculated with the SAP.

into smaller segments, to allow real-time calculations. A brief discussion then follows illustrating a proper choice of the segment size N: if N is chosen too large, nonstationarity becomes an issue. On the other hand, an N too small implies a large number of segments. A reasonable value of N thus must be able to compromise these two aspects. In practice, segments ranging from 10 to 30 ms [2] are used.

The aspect of breaking down the signal by means of a window function is then graphically visualized with emphasis on the distortions introduced by these functions in both time and frequency domains. This discussion is accompanied by figures such as Figure 5.

The experiment ends with a brief comment on the importance of pre-filtering the entire signal to facilitate the speech processing at later stages. The student at this point is able to listen to the speech signal before and after being pre-filtered, to get a real feeling of the true effect of this kind of operation.

Experiment 2: Linear Prediction Analysis

This experiment deals with the linear prediction (LP) analysis where one extracts the so-called LP coefficients. The experiment starts with a segmented signal, as the one resulting from Experiment 1, on which we solve the LP problem, as described in the subsection Linear Prediction Analysis. It is then shown to the student how a small number of LP coefficients can model an entire speech segment (in the order of 80–240 points, as given before), by comparing these two representations in the frequency domain.

The analysis above is repeated for several orders of the LP model, ranging from $M = 5$ (very small) to $M = 30$ (extremely high). The resulting magnitude response of the AR model is depicted for each value of M. It is then shown how the best results of the LP analysis are achieved with M within the range of 8–15. In fact, for small values of M, the AR filter cannot model the speech signal adequately. A large M, however, may increase the number of coefficients beyond necessity.

Experiment 3: Cepstrum Analysis

This experiment deals with the somewhat obscure concept of cepstrum analysis. It starts with an audio explanation pointing out the importance of separating the pitch information from the vocal-tract model, as done, for instance, in the LP analysis, described in Experiment 2.

We then introduce the cepstrum domain, as defined in the subsection Cepstrum Analysis, where a vocal-tract model can be obtained apart from the pitch information.

Following that, it is explained how the cepstrum coefficients can be obtained from the LP coefficients or from the definition of cepstrum followed by a lowpass liftering operation. The computational complexity of these two approaches for determining the cepstrum coefficients is quantified and their overall results are quantitatively compared. Comparisons between the LP and cepstrum analyses are performed and further remarks are added.

CONCLUSION

This paper proposed a new format for a course on speech analysis. The central idea was to combine the endless possibilities for distance learning over the Internet with the interesting research subject of speech signal processing. In that sense, computer experimentation is greatly emphasized, thus helping the student to grasp some concepts such as signal segmentation, linear prediction analysis, and cepstrum analysis.

Two software tools used in the project were presented: the signal-analysis program (SAP) developed at the Federal University of Rio de Janeiro, and the eTEAM, a well-known tool for creating distance-learning class modules. Three experiments were described and made available in the Internet for downloading. Other modules are currently under development and should incorporate all different aspects of speech processing such as coding, synthesis, and recognition.

REFERENCES

[1] T. P. Barnwell III, K. Nayebi, and C. H. Richardson, Speech coding: a computer laboratory textbook, John Wiley & Sons, New York, 1996.

[2] J. R. Deller, J. G. Proakis, and J. L. Hansen, Discrete-time processing of speech signals, Macmillan, New York, 1993.

[3] L. Rabiner and J. Huang, Speech recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[4] A. Antoniou, Digital filters: analysis, design, and applications, 3rd ed., McGraw-Hill, New York, 1996.

BIOGRAPHIES

Vagner L. Latsch was born in Petrópolis, RJ, Brazil, and is currently working towards a BSc degree in electronics engineering at the Universidade Federal do Rio de Janeiro. His research interests include signal processing and distance learning.

Fernando Gil V. Resende, Jr., received his BSc degree from Instituto Militar de Engenharia, Brazil, in 1990, and his MSc and PhD degrees from Tokyo Institute of Technology, Japan, in 1994 and 1997, respectively, all in electrical engineering. Since 1998 he has been with the Department of Electronics Engineering, Universidade Federal do Rio de Janeiro, as an associate professor. His research interests include speech processing and adaptive filtering theory.

Sergio L. Netto was born in Rio de Janeiro, RJ, Brazil, in 1967. He received the BSc degree from the Universidade Federal do Rio de Janeiro (UFRJ) in 1991, the MSc degree from the COPPE/UFRJ in 1992, and the PhD degree from the University of Victoria, Canada, in 1996, all in electrical engineering. Since 1997 he has been with the Department of Electronics Engineering at UFRJ, as an associate professor. Since 1998 he has also been with the Program of Electrical Engineering at COPPE/UFRJ. His teaching and research interests include digital filter design, adaptive IIR filters, and speech signal processing.