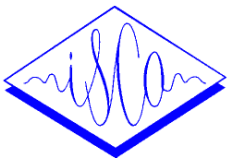

ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education (MATISSE)
University College, London, UK
April 16–17, 1999

Human Speech Production Based on a Linear Predictive Vocoder – An Interactive Tutorial

Klaus Fellbaum[1], *Jörg Richter*[2]

[1] Communication Engineering, Brandenburg Technical University of Cottbus, Germany
fellbaum@kt.tu-cottbus.de , [2] Technical University of Berlin, Germany

Abstract

This tutorial explains the principle of the human speech production with the aid of a Linear Predictive Vocoder (LPC vocoder) and the use of interactive learning procedures. The components of the human speech organ, namely the excitation and the vocal tract parameters, are computed. The components are then fed into the synthesis part of a vocoder which finally generates a synthesised speech signal. The user can replay the signal and compare it with the reference speech signal. For visual comparison, the reference speech signal and the reconstructed speech signal are depicted in both, the time and frequency domain.
For the reconstructed signal, also the pitch frequency contour is graphically presented and the user can directly manipulate this contour. The main advantage of the tutorial are its numerous interactive functions. The tutorial is based on HTML pages and Java applets and can be downloaded from the WWW.

1. Introduction

Speech is human's most important form of communication. This explains why the human speech production has always been of fundamental interest and many attempts have been made to construct speaking machines. Thanks to the progress in electronic speech processing it is now possible to analyse spoken utterances and to extract control parameters which can be used to produce human-like speech with a kind of *artificial mouth*. The main difference to recorded and replayed speech (which can be produced by any tape recorder) is twofold: a) the user has access to speech components and can flexibly manipulate speech characteristics and b) the artificial mouth can produce *any* speech utterances, that means, even utterances which were not spoken before by a human.
It is important to state that the understanding of the human speech production can be optimally used for the development of very efficient schemes for speech coding techniques.
As an illustrative example for a speech coder, based on the principle of an artificial mouth we have selected the vocoder, which will be described in detail in the next section. In the scope of our tutorial, the vocoder has the advantage that the coding principle is directly related to the excitation and articulation function of the human speech organs.
The tutorial is designed for students of various disciplines like communication engineering, physics, linguistics, phonetics, medicine, speech therapy etc... It requires some basic knowledge in signal processing. For example, the student should know how to read and interprete a time signal and a spectrogram.
The tutorial is based on HTML pages and Java applets. It is available from our WWW server, the address is

http://www.kt.tu-cottbus.de/speech-analysis/ .

For the use of the program, the Windows platform is required and we recommend the Netscape browser 4.04 (or higher) or better 4.5. More instruction details are given in the tutorial text.
As to recording the own voice, we had severe problems since Java (up to now) does not support speech input. Fortunately there is the shareware *SoundBite* of the *Scrawl* company, based on *JNI (Java Native Interface)* which offers speech input facilities. It can be taken from our WWW server mentioned above. For the audio output we use *sun.audio*, which is part of common browsers. If there is no need (or interest) to record the own voice and to restrict on the stored speech samples, no shareware is necessary.

2. Human Speech Production

Speech is produced by a cooperation of lungs, glottis (with vocal cords) and articulation tract (mouth and nose cavity). Fig.1 shows a cross section of the human speech organ. For the production of voiced sounds, the lungs press air through the epiglottis, the vocal cords vibrate, they interrupt the air stream and generate a quasi-periodic pressure wave. In the tutorial, a short demonstration of the vibrating vocal cords is presented after clicking on the video symbol under Fig.1.
The pressure impulses are the well-known pitch impulses and the frequency of the pressure signal is the *pitch frequency* or *fundamental frequency*. It is the part of the voice signal that defines the speech melody. When we speak with a constant pitch frequency then the speech sounds monotonous, but in normal cases a permanent change of the frequency ensues.
The pitch impulses stimulate the air in the mouth and for certain sounds (nasals) also the nasal cavity. When the cavities resonate, they radiate a sound wave which is the speech signal.

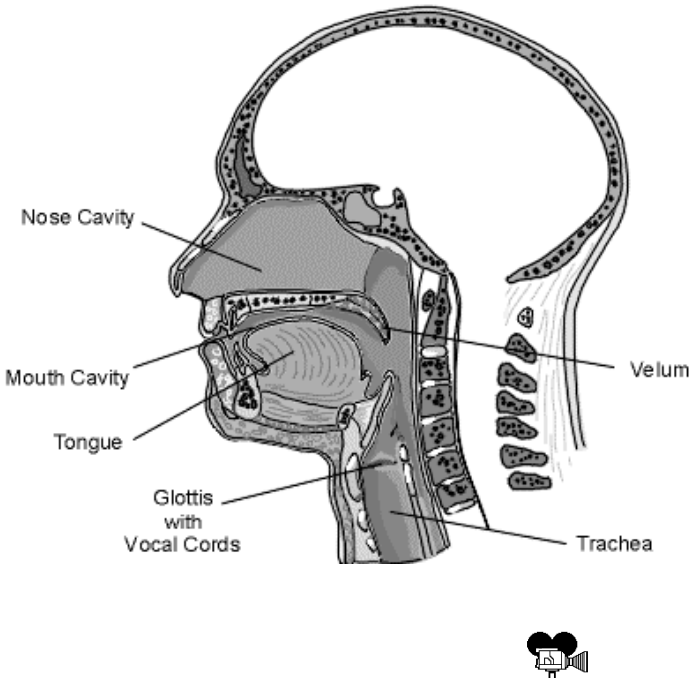

Fig. 1: Human speech production

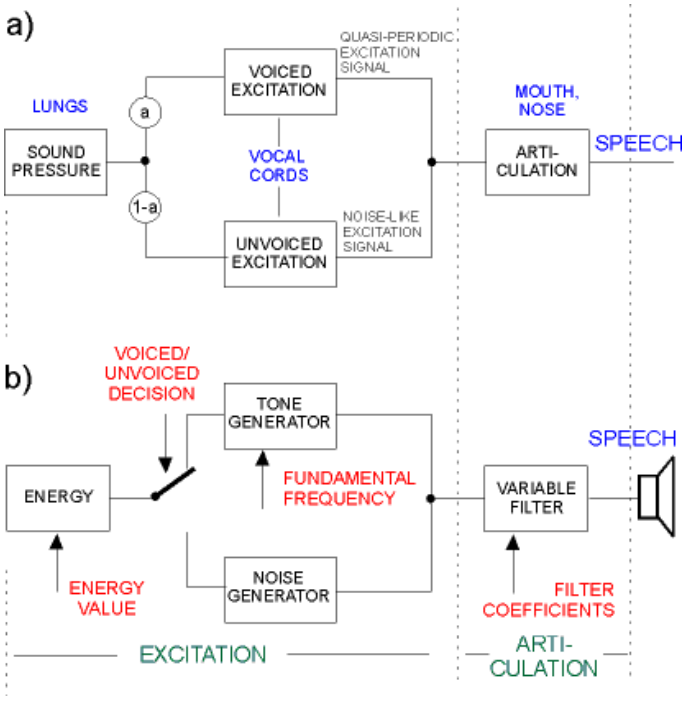

Fig. 2: Models of the human speech production.
a) function model, b) simplified model

Both cavities act as resonators with characteristic resonance frequencies, the mouth resonance frequencies are called *formants*. Since the mouth cavity can be greatly changed, we are able to pronounce very many different sounds.
In the case of *unvoiced* sounds, the vocal cords are open and the excitation of the vocal tract is more noise-like.

3. Speech Production by a Linear Predictive Vocoder

The human speech production can be illustrated by a simple model (Fig.2a) which generates speech according to the mechanism described above. It is important to state that in practice all sounds have a mixed excitation, that means, the excitation consists of voiced and unvoiced portions.
As a matter of fact, the relation of these portions varies strongly with the sound being generated. In our model, the portions are adjusted by two potentiometers [2].

Based on this model, a further simplification can be made (Fig.2b). Instead of the two potentiometers we use a 'hard' switch which only selects between voiced and unvoiced excitation. The filter, representing the human articulation tract, is a simple recursive digital filter; its resonance behaviour (frequency response) is defined by a set of filter coefficients. Since the computation of the coefficients is based on the mathematical optimisation procedure of *Linear Predictive Coding*, they are called *Linear Prediction Coding Coefficients* or *LPC coefficients* and the complete model is the so-called

LPC Vocoder (Vocoder is a concatenation of the terms 'voice' and 'coder').
In practice, the LPC Vocoder is used for speech telephony. A great advantage of it is the very low bit rate needed for speech transmission (in the order of 3 kbit/s, compared to PCM with 64 kbit/s). For more details see [1] and [4].
The main reason why we use the LPC vocoder in our tutorial are the manipulation facilities and the narrow analogy to the human speech production. Since the main parameters of the speech production, namely the pitch and the articulation characteristics, expressed by the tone generator and LPC coefficients, are directly accessible, the audible voice characteristics can be widely influenced. For example, the transformation of a male voice into the voice of a female or a child is very easy; this - among others - will be demonstrated in the tutorial. Also the number of filter coefficients can be varied to influence the sound characteristics, since these coefficients are related to the formant frequencies.

4. Vocoder Simulation

Fig. 3. shows the simulation module of our LPC vocoder as a block diagram. The user can either record his or her own voice via microphone or load samples of prerecorded speech. This speech signal always serves as *reference signal* for further investigations, above all, for acoustic and visual comparisons.

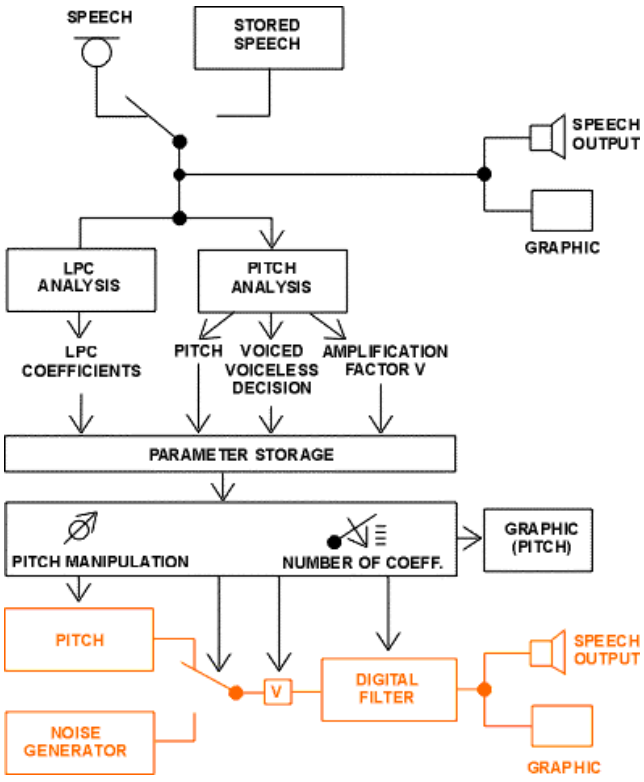

Fig. 3: Scheme of our experimental system

The next steps are the LPC and the pitch analysis. Both, the set of LPC coefficients and the pitch values are then stored in the parameter memory. These parameters are needed to control the synthesis part of the vocoder which is shown in the lower part of the diagram. Obviously, it has the same structure as the model shown in fig. 2b.

The pitch values (pitch contour) and the number of prediction coefficients can be changed and these changes have a significant influence on the reconstructed speech, as mentioned earlier.

We will now describe the different presentation forms, selection procedures and manipulation facilities.

Fig. 4 presents the interactive user interface for the speech processing experiments. The upper diagram displays the reference speech signal. It can be presented as time signal or frequency spectrum (visible speech diagram).

The lower diagram shows the result of the LPC analysis and synthesis. The user can select the speech signal (either the time signal or spectrum) or the pitch sequence as a bar diagram (this is shown in the lower part of fig. 4).

In all display modes each diagram can be scrolled and zoomed and all these manipulations are always applied to both diagrams. Thus the same portion of the speech signal is visible in the upper and the lower diagram.

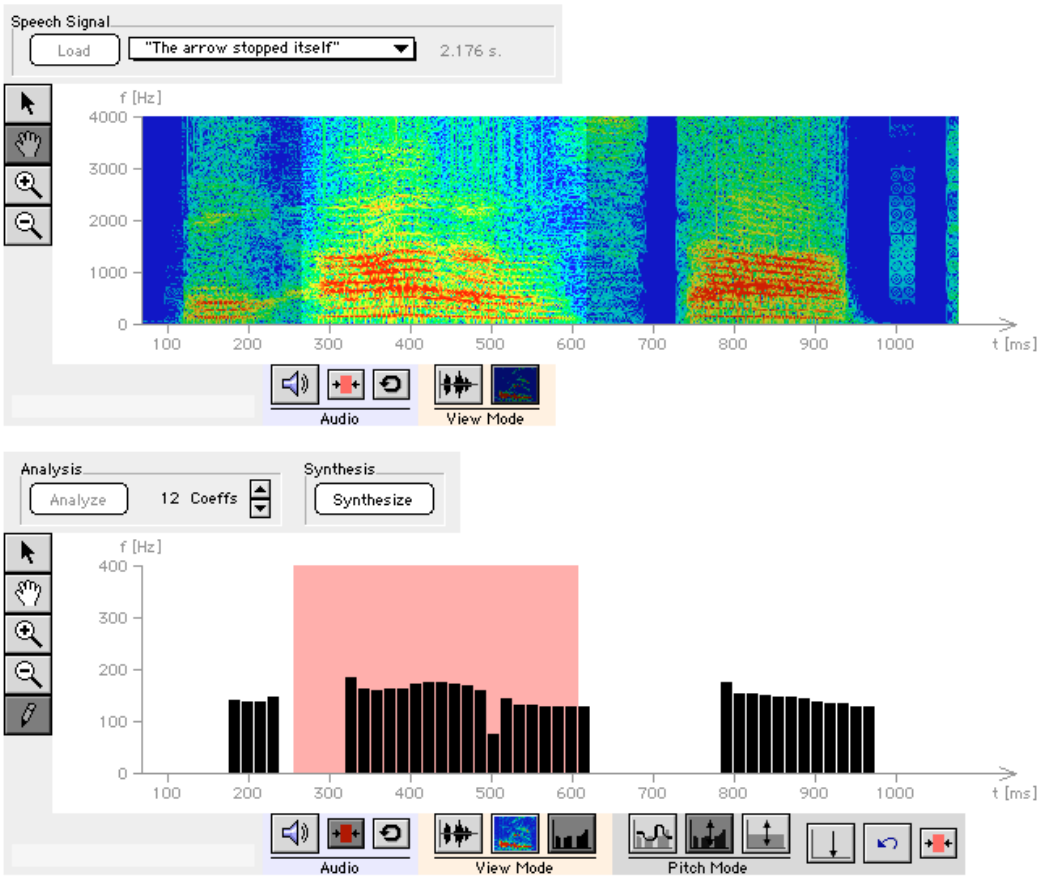

Fig. 4: Interactive user interface of the tutorial

This is very useful for the comparison of the reference speech signal with analysis/synthesis results and the relations between time signal, frequency spectrum and pitch sequence.
Every speech signal can be played back at any time, either as a complete signal or as part of the signal. To set the portion to play, an area of the speech signal can be marked with the mouse. The selected portion is also applied to both diagrams. It is thus easy to explore the relationship between the audible and visual representations of a speech signal.
As to the pitch manipulation, the user has different possibilities, they are controlled by some buttons which are arranged at the bottom of the lower diagram (fig. 4).
For example, all pitch bars can be raised or lowered with a constant value, they can be set to the same value (monotonous speech) including to zero (whisper voice), and each pitch bar can be changed individually which is very important for stress investigations.

5. Concluding Remarks

The tutorial, presented here, was produced with the aim to illustrate the principle of speech production and to arouse interest for the fascinating area of speech communication sciences.
Although the tutorial covers a subject of the electronic speech processing, the main emphasis is put on the visual and acoustical explanation and illustration of the human speech production and on many possibilities to interactively manipulate the speech characteristics. For the user of the tutorial, the best way to get a feeling how manipulations are audible is to start with parameter variations and listening tests in a playful way.
As a very important extension, the tutorial can be a tool for speech therapists. They can record speech disorders and depict them as time signal and spectrum. For comparison, normal speech is shown simultaneously and the deviations are then obvious. Similarily, persons with speech disorders, above all, deaf or hard of hearing persons, who very often have speech organs with full function, get a valuable support when they try to articulate and receive the acoustic result for control as a spectogram, pitch bar diagram or time signal.

Acknowledgements

The tutorial is embedded into the activities of the Socrates/Erasmus Thematic Network "Speech Communication Sciences" and it was funded by the European Network in Language and Speech (ELSNET).

References

[1] Deller, J.R; Proakis, J.G; Hansen, J.H.L. (1993). *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, New York 1993
[2] Fellbaum, K. (1984*). Sprachverarbeitung und Sprachübertragung*, Springer-Verlag, Berlin 1984
[3] Hess, W. (1983). *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin 1983
[4] Jayant, N.S.; Noll, P.(1984*). Digital Coding of Waveforms*, Prentice-Hall, 1984