ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

HUMAN SPEECH PRODUCTION – AN INTERNET-BASED INTERACTIVE MULTIMODAL TUTORIAL

Klaus Fellbaum
Brandenburg Technical University of Cottbus, Germany,
Universitaetsplatz 3-4, D-03044 Cottbus
Joerg Richter
Technical University of Berlin, Germany
fellbaum@kt.tu-cottbus.de

ABSTRACT

We are presenting a tutorial which describes how a human produces speech and how this speech production can be simulated by an LPC vocoder.
In the first section, we describe that speech is generated by a periodic or noise-like signal of the vocal cords which excites the articulation tract (mouth, nose cavity).
The next section explains how this natural sound production is simulated by an LPC vocoder system.
The vocoder was chosen because it gives the user access and manipulation possibilities to the speech parameter, above all, the pitch and the formant structure, expressed by the LPC coefficients.
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.

Keywords: Speech tutorial, LPC-vocoder

10.21437/Eurospeech.1999-532x

1. INTRODUCTION

The investigation of the human speech production is a very fascinating area. In the past, many attempts have been made to construct speaking machines but only poor results occurred. Thanks to modern technology we are now able to produce synthetic speech which sounds rather natural.
A very successful approach is based on the simulation of the human speech organs by a system called VOCODER. The term „vocoder“ is a concatenation of „voice“ and „coder“ which means that the coding procedure has a direct relation to the human voice production. The vocoder principle will be explained in the next section.
The tutorial is designed for students of various disciplines like communication engineering, physics, linguistics, phonetics, medicine, speech therapy a.s.o.. 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 [1].
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 4.06* (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

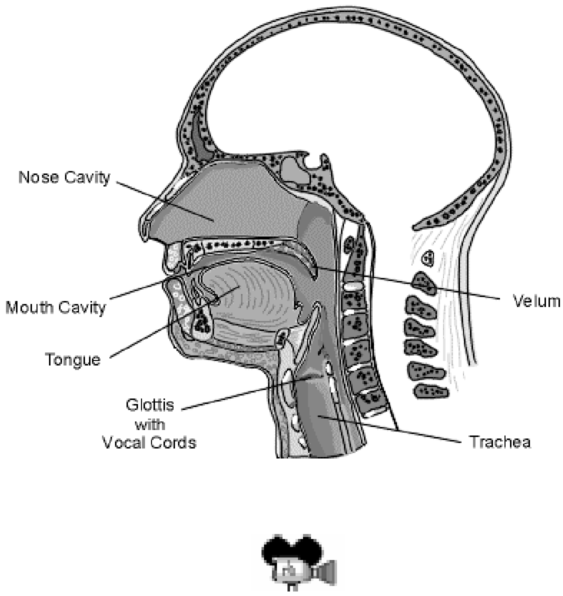

Fig. 1: Human speech production

lungs press air through the epiglottis, the vocal cords vibrate and they produce 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 [3].

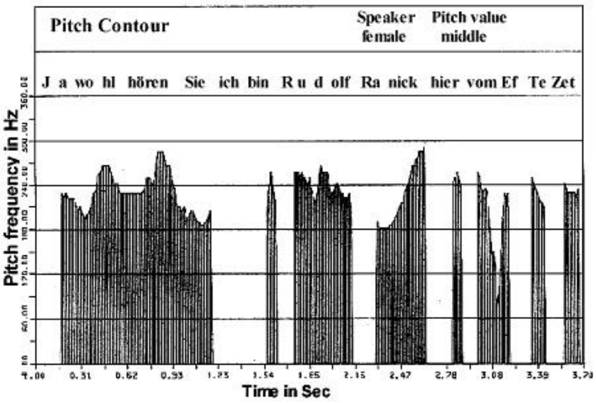

Fig. 2: Pitch sequence of the German sentence „Jawohl hören Sie, ich bin Rudolf Ranick hier vom FTZ“

Fig.2 shows the sequence of pitch impulses for a German sentence. The strong variations (dynamic) of the pitch frequency over time are obvious.
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. Both cavities act as resonators with characteristic resonance frequencies, called *formant frequencies.* 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.

Fig. 3: Positions of the articulation organs for the two sounds „m“ and „t“

As examples, fig. 3 shows the mouth positions when the sounds „m“ and „t“ are pronounced. In the tutorial, they are audible by a mouse click.
The human speech production can be illustrated by a simple model (Fig.4a), 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. Of course, the relation of these portions varies strongly with the sound being generated. In our model, the portions are adjusted by two potentiometers [2]. The articulation tract, which in nature consists of the nose and mouth cavity, is simply replaced by one module.

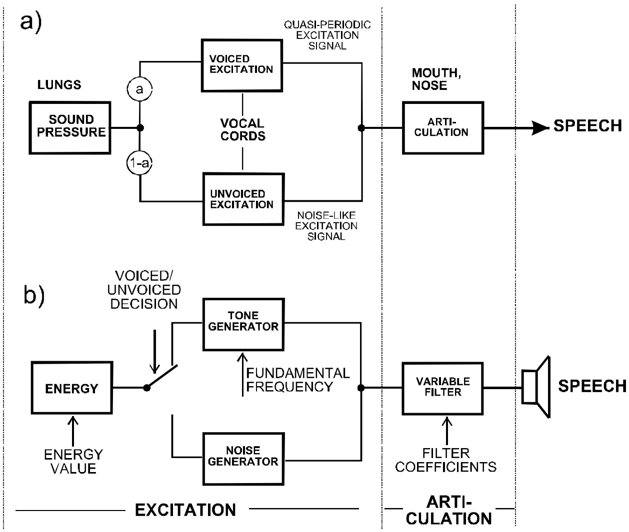

Fig. 4: Models of the human speech production,
a) function model, b) technical realisation by a vocoder

3. SPEECH PRODUCTION USING AN LPC VOCODER

Based on this model, a further simplification can be made (Fig.4b). Instead of the two potentiometers, a vocoder uses a 'hard' switch which only selects between voiced and unvoiced excitation since a correct computation of the voiced and unvoiced portion is very complicated. The filter, representing the articulation tract, is a simple recursive digital filter; its resonance behaviour (frequency response) is defined by a set of filter coefficients, in our case the *Linear Prediction Coding Coefficients* or *LPC coefficients*.
In practice, the LPC Vocoder is used for speech coding purposes. Its great advantage is the very low bit rate needed for speech transmission (about 3 kbit/s) compared to PCM (64 kbit/s). However, for many cases (e.g. telephony), the speech quality produced by the vocoder is at the lower bound of acceptance. For more details see [2] 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 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 - can be demonstrated in the tutorial.

Also the number of filter coefficients can be varied to influence the sound characteristics above all the formant characteristics.

4. INTERACTIVE PART OF THE TUTORIAL

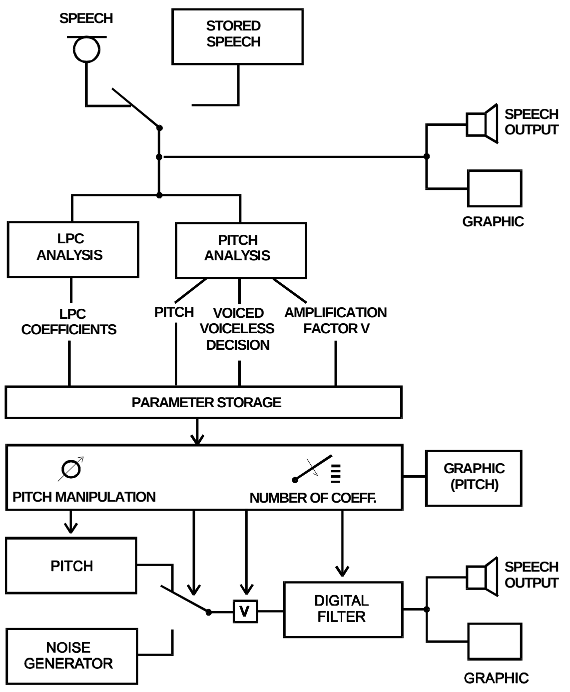

Fig. 5: Scheme of our experimental system

Fig. 5. shows the simulation module of the 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 serves as *reference signal* for further investigations, above all, for acoustic and visual comparisons.

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. 4b.

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. 6 presents the interactive user interface for the speech processing experiments. The upper diagram (Fig. 6a) 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 Fig. 6b). 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. This is very useful for the comparison of the reference speech signal with analysis/synthesis results and the

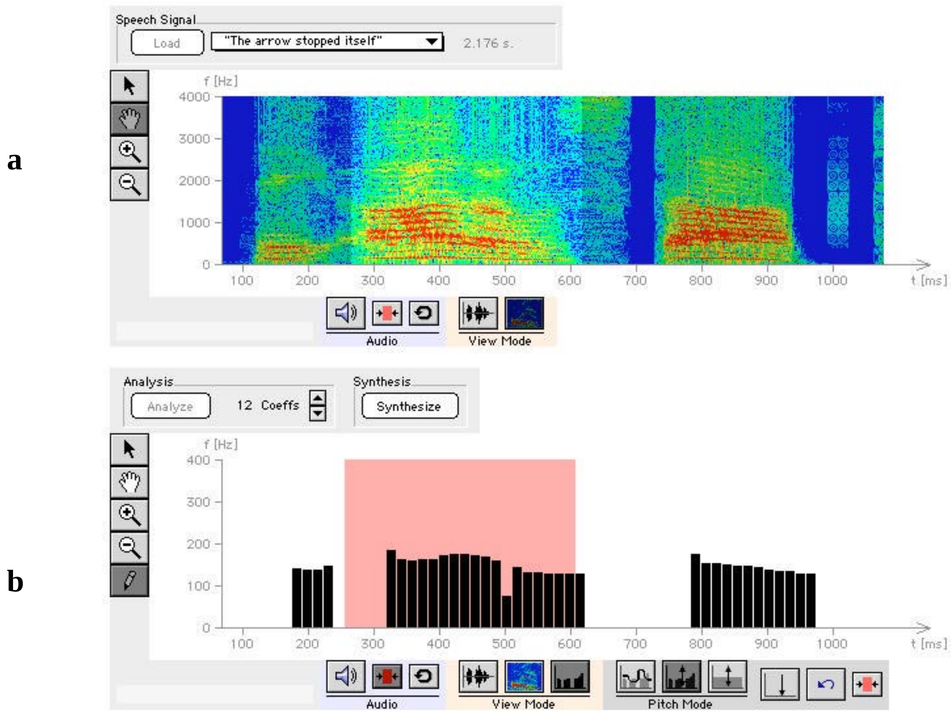

Fig. 6: Interactive user interface of the tutorial

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 is marked by the mouse.
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. 6b). 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 explain which of the speech components influence in which way the resulting speech.
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. It must be emphasized that the user of the tutorial should take the time to experiment with his or her voice in a more playful way and to explore the interrelation between acoustic and visual phenomena of the speech.
In addition, persons with speech disorders, above all, deaf or hard of hearing persons (who very often have speech organs with full functions) have a valuable support when they try to articulate and get the acoustic result for control as a spectogram or time signal.

6. REFERENCES

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

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).