Digital Audio and Speech Processing
(Sayısal Ses ve Konuşma İşleme)

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr
nizamettinaydin@gmail.com
http://www3.yildiz.edu.tr/~naydin

Introduction

Course Outline

- Introduction to Speech and audio processing
- Review of Fundamentals of Digital Signal Processing
- Human hearing mechanism, Speech and audio perception,
- Speech and audio signals modelling, Short term analysis of speech,
- Time domain analysis, Short term Fourier analysis,
- Enhancement of speech and audio, Noise reduction

Course Outline

- Feature extraction of Speech and audio signals
- Midterm 1
- Estimating Speech parameters: Pitch frequency and Formant estimation
- Calculating Mel-frequency cepstral coefficients
- Speech recognition Methods, Vector quantization algorithm
- Automatic Speech Recognition, Hidden Markov Models
- Speech coding and compression methods
- Final

Some Recommended Books

- Speech and Audio Signal Processing: Processing and Perception of Speech and Music, B. Gold, N. Morgan, D. Ellis
- Introduction to Digital Speech Processing, Lawrence R. Rabiner, Ronald W. Schafer
- Theory and Applications of Digital Speech Processing, Lawrence R. Rabiner, Ronald W. Schafer
- Fundamentals of Speech Recognition, Rabiner, L., and Juang, B.-H.
- Discrete-Time Processing of Speech Signals, John R. Deller, Jr., John H. L. Hansen, John G. Proakis
- Digital speech processing, synthesis, and recognition, Sadaoki Furui
- Applied Speech and Audio Processing With MATLAB Examples, Ian McLoughlin
- Audio and Speech Processing with MATLAB, Paul R. Hill
- Audio Processing and Speech Recognition, Soumya Sen, Anjan Dutta, Nilanjan Dey
- ...
- ...

Digital Audio and Speech Processing

- Signal
- Acoustic
- Sound
- Audio
- Voice
- Speech
- Music
- Digital
- Processing

Signal

- A pattern of variations of a physical quantity that can be manipulated, stored, or transmitted by physical process.
- An information variable represented by physical quantity.
- In the mathematical sense it is a function of time, $x(t)$, that carries an information.

Acoustics

- A branch of physics that deals with the study of mechanical waves in gases, liquids, and solids including topics such as vibration, sound, ultrasound and infrasound.
- The application of acoustics is present in almost all aspects of modern society with the most obvious being the audio and noise control industries.
- Science of acoustics spreads across many facets of human society—music, medicine, architecture, industrial production, warfare and more.

7

Sound

- Mechanical radiant energy that is transmitted by longitudinal pressure waves in a material medium (such as air) and is the objective cause of hearing
  - Ultrasound
    - Vibrations of the same physical nature as sound but with frequencies above the range of human hearing
  - Infrasound
    - Vibrations of the same physical nature as sound but with frequencies below the range of human hearing

8

Hearing range in Hertz

	Lowest		Highest
Turtle	20	-	1.000
Goldfish	100	-	2.000
Frog	100	-	3.000
Pigeon	200	-	10.000
Sparrow	250	-	12.000
Human	20	-	20.000
Chinpanzee	100	-	20.000
Rabbit	300	-	45.000
Dog	50	-	46.000
Cat	30	-	50.000
Guinea Pig	150	-	50.000
Rat	1.000	-	60.000
Mouse	1.000	-	100.000
Bat	3.000	-	120.000
Dolphin	1.000	-	130.000

9

Audio

- : of or relating to acoustic, mechanical, or electrical frequencies corresponding to normally audible sound waves which are of frequencies approximately from 15 to 20,000 hertz
- : of or relating to sound or its reproduction and especially high-fidelity reproduction
- : relating to or used in the transmission or reception of sound
- : of, relating to, or utilizing recorded sound

10

Voice

- Sound uttered by the mouth, especially by human beings in speech or song; sound thus uttered considered as possessing some special quality or character
- Sound made through vibration of the vocal cords; sonant, or intonated, utterance; tone;
  - distinguished from mere breath sound as heard in whispering and voiceless consonants.
- The tone or sound emitted by an object

11

Speech

- Human vocal communication using language.
- The physical production of sound using our tongue, lips, palate and respiratory system to communicate ideas.
- Each language uses phonetic combinations of vowel and consonant sounds that form the sound of its words, and using those words in their semantic character as words in the lexicon of a language according to the syntactic constraints that govern lexical words' function in a sentence.

12

Music

- Vocal or instrumental sounds (or both) combined in such a way as to produce beauty of form, harmony, and expression of emotion.
- A form of art that uses sound organized in time.
- A form of entertainment that puts sounds together in a way that people like, find interesting or dance to.
  - Most music includes people singing with their voices or playing musical instruments, such as the piano, guitar, drums or violin
- The written or printed signs representing vocal or instrumental sound.

13

Historical Background

- A brief history of synthetic audio,
  - 8th Century mechanical devices
- Speech and music machines from the first half of the 20th Century
- Systems for analysis and synthesis
- Speech recognition
  - a 20th Century invention

14

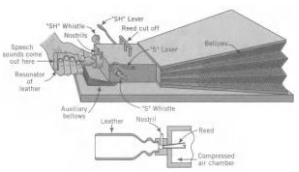
Synthetic Audio

- Wolfgang von Kempelen demonstrated that the speech-production system of the human being could be modeled (1780).
  - He showed this by building a mechanical contrivance that "talked."
    - https://www.youtube.com/watch?v=k_YUB_S6Gpo
    - Dudley, H., and Tarnoczy, T. H., "The speaking machine of Wolfgang von Kempelen," J. Acoust. Soc. Am. 22: 151-166, 1950.
      - https://pdfslide.net/documents/the-speaking-machine-of-wolfgang-von-kempelen-homer-dudley-and-t-h-tarnoczy.html
- Von Kempelen also wrote a book that dealt with the origin of speech, the human speech-production system, and his speaking machine
  - Von Kempelen, W., Le Mechanisme de lapavola, suivi de la Description d'une machine parlante. Vienna: J.V. Degen, 1791.

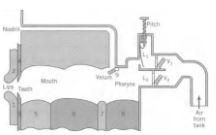
15

Synthetic Audio

- the speaking machine built by Wheatstone that was based on von Kempelen's work.



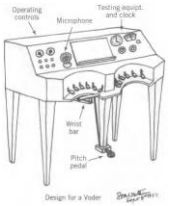
- Riesz's speaking machine



16

The Voder

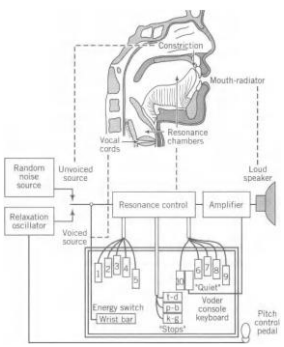
- Modern methods of speech processing began in with the development of two devices.
  - The channel vocoder (voice coder)
  - The Voder (voice-operated demonstrator)



- Dudley, H., Riesz, R., and Watkins, S., "A synthetic speaker," J. Franklin Inst. 227: 739, 1939.
- Sketch of the Voder.

17

Voder controls



- Many speech-synthesis devices were built in the decades following the invention of the Voder, but the underlying principle has remained quite fixed.
  - For many cases, there is a separation of source and filter followed by the parameterization of each.
- The same underlying principles control the design of most music synthesizers.

18

Music Machines

- 17th Century drawing of a water-powered barrel organ
- Spring-powered barrel organs may have existed as long ago as the 12th Century.
  - Barrel organs work on the same concepts as present-day music boxes
    - The barrel organ is a form of read-only memory, and not a very compact form at that.
    - Barrel organs could not record music played by a performer

19

Music Machines

- In the late 18th Century, both of these problems were overcome by melography, which allowed music to be both recorded and played back, using the medium of punched paper tape or cards.
- The idea originated for the automation of weaving and was developed fully by Joseph Marie Jacquard, who designed a device that could advance and register cards.
  - Punched cards were used by Babbage in the design of his computing machine and, in our time, were used by many computer manufacturers such as IBM.

20

Music Machines

- A modern example of a player piano is the solenoid-controlled Bosendorfer at the MIT Media Laboratory.
  - Using this system, Fu [Fu, A. C, "Resynthesis of acoustic piano recordings," M.S. Thesis, Massachusetts Institute of Technology, 1996.] synthesized a Bosendorfer version from an old piano roll by Rachmaninoff.
- At the beginning of the 20th century, a mighty device called the telharmonium was constructed by Thaddeus Cahill.
  - Cahill had the ingenuity to realize that any sound could be synthesized by the summation of suitably weighted sinusoids.
    - https://www.youtube.com/watch?v=x8wn_gJD8kw
    - https://www.youtube.com/watch?v=EG1C0E7L8wU

21

Music Machines

- He implemented each sinusoid by actuating a generator.
- To create interesting music, many such generators (plus much additional equipment) were needed, so the result was a monster, weighing many tons.
- Cahill's concept of additive synthesis is still an important feature of much of the work in electronic music synthesis.
- This is in contrast to many later music synthesizers that employ subtractive synthesis, in which adaptive filtering of a wideband excitation function generates the sound.
  - The additive synthesis concept was used by McCaulay and Quatieri [McAulay, R. J., and Quatieri, T. R, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Signal Process.* 34(4): 744-754, Aug. 1986.] to design and build a speech-analysis-synthesis system.

22

Music Machines

- Another, complete synthesizer is the theremin, named after its inventor, the Russian Lev Termin.
- In this system, an antenna is a component of an electronic oscillator circuit; moving one's arm near the antenna changes the oscillator frequency by changing the capacitance of the circuit, and this variable frequency is mixed with a fixed-frequency oscillator to produce an audio tone whose frequency can be varied by arm motion.
- Thus the theremin generates a nearly sinusoidal sound but with a variable frequency that can produce pitch perceptions that don't exist in any standard musical scale.
  - https://www.youtube.com/watch?v=YNoR-SR5t1s
  - https://www.youtube.com/watch?v=scr7-vTX19I

23

Speech Communication - Background

- Transmission of Acoustic Signals
  - The earliest network for speech communication at long distances was a system called the "stentorian network," which was used by the ancient Greeks.
    - It consisted of towers and men with very loud voices.
- Acoustical Telegraphy before Morse Code
- Using musical instrument to transmit news long distances
- Transmitting sound underwater.
- …

24

Speech Communication - Background

Later a group of inventors, among whom we find Kircher(1601–1680), Scheventer (1636) and the two Bernoulli brothers, sought to transmit news long distances by means of musical instruments each note representing a letter. One of the Bernoullis devised an instrument, composed of five bells, which permitted the principal letters of the alphabet to be transmitted.

It is told that the King of England was able to hear news transmitted 1.5 English miles to him by means of a trumpet. He had this trumpet taken to Deal Castle, whose commander said that this instrument permitted a person to make himself understood over a distance of three nautical miles. It was invented by the "genial mechanic" of Hammersmith, Sir Samuel Morland (1626–1696). It's [sic] mouthpiece was designed so that no sound could escape from either end. Morland published a treatise on this instrument entitled "Tube Stentorophonica" and in 1666 he wrote a report on "a new cryptographic process."

In 1762 Benjamin Franklin experimented with transmitting sound under water. In 1785 Gauthoy and Biot transmitted words through pipes for a distance of 395 meters. But at a distance of 951 meters speech was no longer intelligible.

We can also regard the ringing of bells as acoustical telegraphy or telephony, if we consider that in certain Swiss villages the inhabitants recognize from their tone whether the person who has just died is a man or a woman, a member of a religious order, etc. Moreover, every Sunday the inhabitants of these villages follow the principal passages of the divine service with the aid of the pealing of the different bells. We have seen old people, prevented from attending the service because of their infirmities, with prayer book in hand, follow at a distance the priest's various movements.

Our story would be incomplete if we did not mention the African tom-tom, which some people consider a sort of acoustical telegraphy. The African explorer, Dr. A. R. Lindt, has written a short report on the tom-tom. We quote the following from his work: "There is no key to the acoustical telegraphy of the Africans. Since they have no written language, they are unable to divide their words into letters. The tom-tom therefore does not translate letter by letter or even word by word, but translates a series of well-defined thoughts into signals. There are different signals for all acts

25

Speech Communication - Background

- The Telephone
  - invented by Alexander Graham Bell
    - Bell's primary profession was that of a speech scientist who had a keen understanding of how the human vocal apparatus worked
    - Bell understood the rudiments of the speech spectral envelope.
  - Telephone technology has been mostly concerned with transmission methods.
  - Recently, however, with the growing use of cellular phones in which transmission rate is limited by nature, efficient methods of speech coding have become an increasingly important component of speech research at many laboratories.

26

Speech Communication - Background

- The Channel Vocoder and Bandwidth Compression
  - The channel vocoder [Dudley, H., "The vocoder," Bell Labs Record 17: 122-126, 1939.] was the first analysis-synthesis system.
  - The vocoder analyzer derived slowly varying parameters for both the excitation function and the spectral envelope

- Colton, F. Â., "The miracle of talking by telephone," National Geographic 70(4): 395-433, 1937.
- Bennett, W. R., "Secret telephony as a historical example of spread-spectrum communication," IEEE Trans. Commun. COM-31: 98-104, 1983.

- Lower Broadway in 1887.

27

Voice-Coding Concepts

- To understand why a device such as a vocoder reduces the information content of speech, we need to know enough about human speech production to be able to model it approximately.
- Then we must convince ourselves that the parameters of the model vary sufficiently slowly to permit efficient transmission.
- Finally, we must be able to separate the parameters so that each one is coded optimally.
- The implementation of these concepts is captured by the phrase "analysis-synthesis system."
- The analysis establishes the parameters of the model; these parameters are transmitted to the receiving end of the system and used to control a synthesizer with the goal of reproducing the original utterance as faithfully as possible.

28

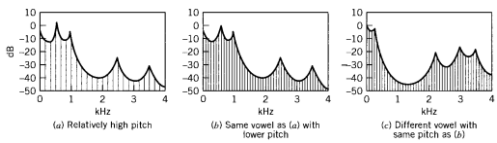
Voice-Coding Concepts

- The speech wave is the response of the vocal tract to one or more excitation signals.
  - This concept leads directly to engineering methods to separate the source (the excitation signal) from the filter (the time-varying vocal tract).
  - The procedures for implementing this separation can be called deconvolution, thus implying that the speech wave is a linear convolution of source and filter.
- In spectral terms, this means that the speech spectrum can be treated as the product of an excitation spectrum and a vocal tract spectrum.

29

Voice-Coding Concepts

- A simplified illustration of the spectral cross section for sustained vowels:



(a) Relatively high pitch (b) Same vowel as (a) with lower pitch (c) Different vowel with same pitch as (b)

- Numerous experiments have shown that such waveforms are quite periodic;
  - this is represented in the figures by the lines.
    - In (a) the lines are farther apart, representing a higher pitched sound;
    - in (b) and (c) the fundamental frequency is lower.

30