Speech Signal Processing – introduction

Jan Černocký, František Grézl
{cernocky,grezl}@fit.vutbr.cz

DCGM FIT BUT Brno

FIT BUT Brno
Agenda

- Who is teaching
- Program of the course
- Assessment
- Literature and Web

- Disciplines connected to the speech processing
- Information content of speech
- Application areas + demonstrations
Who will be teaching

- lectures: Frantisek Grezl, Honza Cernocky
- labs: Frantisek Grezl
- email: grezl@fit.vutbr.cz
Program of the course

- **Lectures** – 12 lectures, 2 hours each
- **Computer labs** – Matlab under Linux
- **Numerical exercise** – examples during the lectures
- **Projects** – there will be one project, given about the half term exam
- will be off most of the may, so will try to finish the course till end of april.
P1 Organization of the course, applications, disciplines.


P3 Speech preprocessing: mean value, preemphasis, frames, basic parameters, spectrogram.
Speech production: vocal apparatus and its model - vocal cords and vocal tract vs. excitation and modification (filter). Basic characteristics in time and spectral domain: influence of the excitation and modification. Formants. Advantages of the
short-term and long-term spectrogram. How to separate excitation and modification: cepstrum + MFCC.
P4 Linear-predictive model: estimating characteristics of the vocal tract, not the excitation ⇒ application in coding and recognition. Prediction of the sample from the previous samples - linear prediction (LP). LP error. Estimation of the LP error. Prediction of the articulation apparatus configuration using LP analysis. Spectrum by means of linear prediction. Parameters derived from LP - LAR and LSF, which are used in encoders. LPC-cepstrum.


P6 Speech encoding I: Objectives. Bit rate, objective a subjective quality measure. Subdivision based on bit rate and quality.
Encoding in time domain. Vocoders - LPC. Vector quantization in speech encoding.

P7 Speech encoding II. - application of CELP, encoding in GSM: GSM, GSM-EFR, GSM-HR, Voice over IP.


P10 HMM II. Continuous speech with a big vocabulary: recognition based on smaller tokens - phnemes... Phonetic structure of a language. Vowels and consonants, characteristics and classification of phonemes. International standards for phoneme denotation: IPA
a SAMPA, TIMIT. Coarticulation.

Recognition: context-dep. tri-phones. Large vocabulary, language modeling, lattice re-scoring, forced alignment.
P11 Recognition features. Requirements - pitch discard, decorrelation. Spectral envelope. What we can do and how we can use it: LPCC, MFCC to decorrelate the features; PCA, LDA, HLDA, normalizations to reduce the influence of the channel. Other improvements - delta, delta-delta. “Hot-topics in parameterization”: TRAPs and FeatureNet, neural nets.

Tools in speech processing.


Generation of the signal in time and frequency domain. Methods of PSOLA a HNM. Applications. SW used in synthesis: EPOS, MBROLA, Festival.
– Speech processing in Matlab.
– LPC and vector quantization, LPT error.
– DTW and HMM in Matlab.
– ? HMM using HTK. most likeli to split DTW and HMM into separate labs.
## Evaluation of the Course

<table>
<thead>
<tr>
<th>category</th>
<th>points</th>
</tr>
</thead>
<tbody>
<tr>
<td>project</td>
<td>29</td>
</tr>
<tr>
<td>midterm - theoretical questions only</td>
<td>20</td>
</tr>
<tr>
<td>final exam - theory and numerical tasks</td>
<td>51</td>
</tr>
<tr>
<td>altogether make</td>
<td>100</td>
</tr>
</tbody>
</table>

- Materials are **NOT** allowed during exams
LITERATURE

- Gold B., Morgan. N.: Speech and audio signal processing, John Wiley & Sons, 2000 [library]
INTRODUCTION INTO AUTOMATIC SPEECH PROCESSING

Definition

“Automatic speech processing allows voice communication between people (encoding) or between a human being and a machine.”
Disciplines involved in Speech Processing

Speech processing is a pluri-disciplinary field, utilizing knowledge of natural sciences, technical sciences and social sciences.

- **Physiology:** study of articulatory and auditory apparatus, Knowledge used to facilitate design of the model.
- **Acoustics:** studies physical mechanisms of production and perception of speech.
- **Signal processing:** multiple areas: modeling, parameterization, identification, spectral analysis, encoding, informational theory, pattern recognition, etc.
- **Social sciences**
  * *phonetics* – a subfield of linguistics that comprises the study of the sounds
of human speech.

* **phonology** – s subfield of linguistics that deals with the sound systems of languages. Subdivides speech into basic units, phonemes. *Phoneme* is the smallest posited structural unit that distinguishes meaning, though they carry no semantic content themselves.

* **prosody** – a study of the sound of the language (melody, duration of phonemes, accent in words and sentences, ...).
* **lexicology** – is that part of linguistics which studies words, their nature and meaning, words’ elements, relations between words (semantical relations), word groups and the whole lexicon.
* **grammar** – is the field of linguistics that covers the conventions governing the use of any given natural language. Grammar is the set of rules describing use of the language. Important for the synthesis.
* **syntaxis** – is the study of the principles and rules for constructing sentences in natural languages.
* **semantics** – is the study of meaning in communication. Basic item is usually a word.
the goal is to estimate *informational speed* (in bits per second, bit/s or bps), required to express speech in different formats. For comparison, we will express phonetic and acoustic form in digital representation.

**Phonetic Form**

number of phonemes in Czech is 36. To calculate *informational quantity* we will consider a source generating mutually independent elements $x_i$ from the set $X = \{x_1, \ldots, x_S\}$, where $S$ is a finite set of items. Each item has the probability $p(x_i)$ and the items constitute a complete system, thus:

$$\sum_{i=1}^{S} p(x_i) = 1.$$  (1)
Informational content of the $i$-th item is given by the number of bits we need to express the item:

$$I(x_i) = - \log_2 p(x_i) \quad [\text{bit}].$$

(2)

Source entropy (average information value) is given by:

$$H(X) = - \sum_{i=1}^{S} p(x_i) \log_2 p(x_i).$$

(3)

For the Check phonemes, in case we assume the same probability of occurrence the entropy is $H(X) = 5.2$ bits. Considering the true phonemes probability, the entropy becomes $H(X) = 4.6$ bits.

Considering the mutual dependency (conditional probability) of the phonemes (bi-grams: $p(x_j|x_i)$, trigrams: $p(x_k|x_i x_j)$, etc.), the value of the entropy becomes $H(X) = 3–3.5$ bits.

In the average Czech conversation, a human being produces cca
80–130 words per minute, thus about 10 phones per second. Informational speech is then approximately $C_{phn} = 30–40$ bit/s. Psychoacoustic tests show, that a human being is able to process incoming information with the speed of app. 50 bit/s.
In case speech is represented by a digital signal, the Nyquist–Shannon–Kotelnikov theorem must hold:

\[ F_s > 2F_m, \]

where \( F_s \) is the sampling frequency and \( F_m \) is the highest frequency presented in the spectrum of the signal. Each sample is quantized by one of the \( m \) of allowed quantization levels, which can be expressed by \( N = \log_2 m \) bits. Signal to noise ratio is proportional to \( 6N \) (in dB). The informational speed can be hence expressed as:

\[ C_{ak} = \frac{I}{t} = \frac{\log_2 m}{T_s} = NF_s. \]
Examples:

1. We are given a signal in Hi-Fi quality, $F_s=44.1$ kHz, $N=16$ bit. Resulting informational speed is $C_{ak} = 705$ kbit/s.

2. For a signal in telephone quality (with the band from 300 to 3400 Hz): $F_s=8$ kHz, $N=8$ bit. Resulting informational speed is $C_{ak} = 64$ kbit/s.

Conclusion

We can see from the example that the acoustic form comparing to the phonetic form is greatly *redundant*. Along to the information contained in the phonetic form, the acoustic form carries information about the speaker, their mood, environment and so on. The listener then subconsciously separates different informations in their brain. Unfortunately, it is not know so far how exactly. Nevertheless, it is useful
to utilize the basic knowledge about the speech production to *lower the informational speed*. 
APPLICATION DOMAINS IN SPEECH PROCESSING

Encoding:

facilitates transfer and storage.

**Goal:** represent the speech on the smallest possible number of bits.

**Requirements:**

- complexity ↓,
- delay ↓,
- intelligibility ↑,
- natural sound ↑,
- robustness against errors in the channel ↑.
Speech Signal Processing – introduction. Frantisek Grezl, DCGM FIT BUT Brno 26/32
Demo — ©Andreas Spanias (Arizona University)
http://www.eas.asu.edu/~speech/table.html
• of speech (Speech Recognition)
  – isolated words
  – continuous words (for instance, figures in telephone number credit cards).
  – continuous speech – the hardest task, still not working in all cases, especially in cases with large vocabulary (LVCSR - Large Vocabulary Continuous Speech Recognition).

• of the speaker (Speaker Recognition)
  – identification – who is the speaker given a set of reference speakers?
  – verification – are two speech segments coming from the same speaker?
What we have?

- Spontaneous Speech
  - word spotting
  - system driven dialog
  - form fill by voice
  - voice commands

- Fluent Speech
  - digit strings
  - name dialing
  - directory assistance

- Read Speech
  - speaker verification

- Connected Speech
  - natural conversation
  - 2-way dialog
  - network agent & intelligent messaging

- Isolated Word
  - transcription
  - office dictation

Vocabulary Size: 2, 20, 200, 2000, 20000

synthesis:

computer has to generate speech it never “heard” before (e.g. the speech wasn’t recorded from a human speaker). The most difficult is synthesis from text. (TTS – text-to-speech).

demos:
http://www.cs.indiana.edu/rhythmmsp/ASA/Contents.html
Other applications...

- medicine: examination of the abnormality and illnesses of the vocal tract.
- psychology a criminalistics: lie detector, estimation of the level of alcohol in blood, stress detection...
- aiding of handicapped (helping improving pronunciation to the deaf etc.)
- language identification
- keyword spotting