Series Course, Summer Term 501130
School of Engineering, The University of Tokyo

Applied Acoustics

Introduction : Presentation of the Course

Shigeki SAGAYAMA
Department of Mathematical Engineering and Information Physics,
School of Engineering, The University of Tokyo
sagayama@hil.t.u-tokyo.ac.jp

Location of the material: http://hil.t.u-tokyo.co.jp/~sagayama/applied-acoustics/

(Notice: This material is a translation of the original material in Japanese – translation errors may still exist)

The Aims of Applied Acoustics

- Aim and Nature
  - Can be considered as ‘Signal Processing Theory III’, following ‘Signal Processing Theory I, II'.
  - Statistical Signal Processing, Pattern Processing, Non-stationary time series modeling
  - Probabilistic Modeling, Statistical Training
- Object domains:
  - Phonetics and Speech Analysis
  - Speech Coding
  - Speech Recognition
  - Speech Synthesis
- Main points of the course
  - Comprehension of the basic algorithms
  - Fundamentals of Statistical Signal Processing
- Prerequisite
  - Theory of Linear System and Fourier Analysis
  - Mathematical Statistics (Distribution, Likelihood)

Prerequisites, Evaluation, General precautions

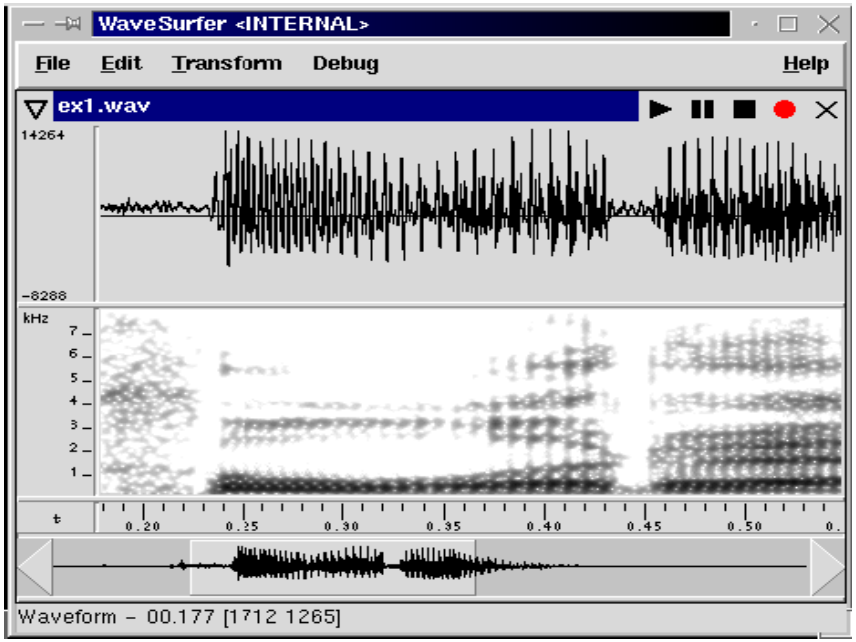
- Prerequisites
  - Statistical Signal Processing (Fourier analysis, spectrum estimation, sampling, etc...)
  - Mathematical Statistics (random variables, distribution, Likelihood, etc...)
- Evaluation
  - Evaluation of the attitude during the class (enthusiasm, attendance, seating position, participation in Q&A, etc.)
  - Score on the final exam
- General Cautions
  - Chatting is strongly prohibited
  - Should be seated in the front part of the class
  - Clothes, Behavior

Topics addressed in the 4th year 'Applied Acoustics' course of the School of Engineering (Provisional)

- What is speech ?
  - Phonetics of the Japanese language
  - (phonology, English and Japanese vowels and consonants, syllables, prosody, why are Japanese people bad at English?)
- Short-time spectrum analysis
  - Sampling theorem, Quantization noise
  - Preemphasis, window of wave pattern
  - Short-time autocorrelation function
  - Short-time spectrum analysis, pitch structure
  - Short time cepstrum analysis
- All-pole models
  - Linear models
  - Linear Predictive Coding (LPC)
  - EM algorithm and HMM* Training
  - Residual signal, pitch extraction
  - Partial Autocorrelation (PARCOR)
  - Line Spectrum Pair (LSP)
- Spectral distance measures
  - Itakura/Saito distance
  - Euclidean distance, Mahalanobis distance
- Clustering analysis
  - *k*-means clustering
  - Scalar Quantization and Vector quantization
- Gaussian Mixture distribution and EM algorithm
  - Non linear time warping
- Dynamic Time Warping (DTW), DP Matching
  - One Pass DP algorithm
- Probabilistic models for speech
  - Multi-dimensional gaussian distribution,
  - Multi-dimensional gaussian mixture distribution
  - Markov process
- Hidden Markov Models (HMM)
  - Viterbi algorithm
  - Training of probabilistic models
  - Viterbi training (Time axis clustering)
  - Embedded training*
- Speech recognition systems
  - Spoken word recognition
  - Large vocabulary speech recognition,
- Continuous speech recognition

Example 1: Useful tools to analyze speech

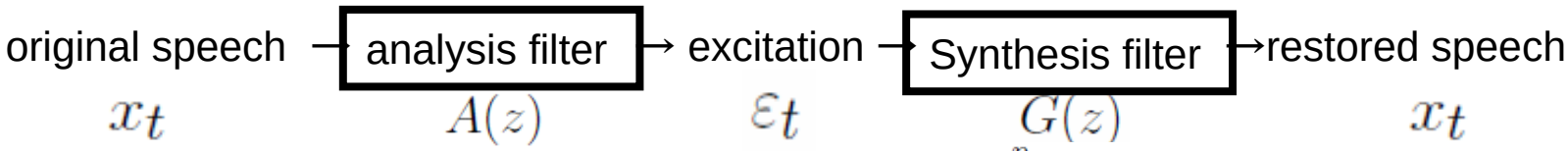
- wavesurfer (KTH)
  http://www.speech.kth.se/wavesurfer/



Wavesurfer screen
works on Linux,
Windows (demonstration)

- spwave (Written by Banno, Itakura Laboratory, the University of Nagoya (currently with Meijo University))
  http://www.itakura.nuee.nagoya-u.ac.jp/people/banno/spLibs/spwave/index-j.html

Example 2: LPC Speech Analysis-Synthesis Model



- Analysis filter: All-zero analysis filter $A(z) = \sum_{i=0}^{p} a_i z^{-i}$
  - Impulse response: $\{1, a_1, a_2, a_3, \cdots . a_n . 0 . 0 . \cdots . \}$
- Synthetic filter: All-pole synthesis filter $G(z) = 1/A(z) = \frac{1}{\sum_{i=0}^{p} a_i z^{-i}}$
  - Speech production model filter
- Excitation signal $\varepsilon(t)$
  - Using the excitation $\varepsilon(t)$ obtained from LPC analysis - reproduction of speech
  - Dividing the excitation $\varepsilon(t)$ into white noise (unvoiced sounds) / impulse train (voiced sounds) - speech analysis-synthesis
  - Approximating the excitation $\varepsilon(t)$ with impulse trains - multipulse coding
  - Creating a codebook of the residual signal of the excitation $\varepsilon(t)$ - CELP (Today's mainstream)
  - Replacing the excitation $\varepsilon(t)$ by another signal - speech conversion (ex. 1, ex. 2)

Example 3: Vector Quantization - basic theory of efficient speech coding

In compressing the speech (as well as other kinds of) information, we need a technique to replace vector volume with signal. Such algorithm is called 'vector quantization' and makes use of training algorithms based on the clustering algorithm.

- *k*-means clustering algorithm
- LBG algorithm
- Segmental k-means algorithm

Example 4: Speech Recognition - non linear time warping

In speech recognition, the problem is how to model the temporal and spectral fluctuations, train the models, and match them to the observation.

The two methods commonly applied are:

- DP (Dynamic Programming) matching algorithm
- What is HMM (Hidden Markov Model) ?

Keywords: Items to be well understood 1/3

- A: What is speech ?
  - Distinction between phonetics and phonology, vowels (three elements of vowels), consonants (three elements of consonants)
  - Phonotactic constraints, syllables, morae, syllabic structure in Japanese (doubled consonants, nasal sounds, long vowels), open syllables
  - 'Rendaku' phenomena, devocalizaion, context dependency of phoneme modification
  - Accentuation in Japanese, what is *m*-mora type-*n* accent, accent combination, accent phrase
  - Why are Japanese people bad at English ? – Hypothesis (phonetic differences between Japanese and English)
- B: Short-time spectral analysis
  - Sampling theorem, Nyquist frequency
  - Quantized noise, its po $\frac{q^2}{12}$
  - Preemphasis, windowi.g
  - Short-time autocorrelation function, positive definite Toeplitz matrix, Wiener-Khinchine theorem, Herglotz theorem
  - Periodogram, Short-time spectrum analysis, pitch structure, spectral envelope
  - Cepstrum, short-time cepstral analysis, pitch estimation (cepstrum method)

Keywords: Items to be well understood 2/3

- C: all-pole model (LPC, PARCOR, LSP)
  - Linear models for speech production, principle of the speech analysis-synthesis
  - All-pole models, stability of recursive digital filter, Schur-Cohn condition
  - Linear Predictive Coding (LPC), normal equation (Yule-Walker equation)
  - Residual signal, pitch extraction (residual correlation method)
  - Partial Autocorrelation Analysis (PARCOR), PARCOR coefficient
  - Levinson-Durbin algorithm
  - Lattice filter, all-pass digital filter, acoustic tube model
  - Line Spectrum Pair (LSP), LSP frequencies
  - Orthogonal polynomial system weighed by the speech spectrum
- D: Spectral distance measures and Clustering analysis
  - Feature vector space, Euclidean distance, Mahalanobis distance, Itakura-Saito distance
  - Clustering, *k*-means algorithm, Convergence (toward the minimum value)
  - Scalar quantization and vector quantization, speech coding･compression･transmission
  - Gaussian distribution, multi-dimensional Gaussian distribution, multi-dimensional Gaussian mixture distribution
  - Maximum likelihood, Kullback-Leibler divergence
  - Incomplete data problem, EM algorithm

Keywords: Items to be well understood 3/3

- E: Non linear time warping
  - Dynamic Time Warping (DTW), DP matching, alignment path, alignment window, traceback
  - Two-stage DP matching, Level Building algorithm, one-pass DP algorithm
  - Segmental *k*-means algorithm
- F: Probabilistic models for speech, training
  - Stochastic stationary signal source, multi-dimensional Gaussian mixture distribution
  - Markov processes
  - Hidden Markov Models (HMM), Mealy type/Moore type, trellis
  - Viterbi algorithm, Viterbi path, Viterbi alignment, Viterbi training (time axis clustering)
  - Forward algorithm, Baum-Welch algorithm, EM algorithm
  - HMM training, embedded training
- G: Speech recognition system
  - Isolated-word speech recognition
  - Language model (network grammar, *n*-gram grammar)
  - Large vocabulary speech recognition, Continuous speech recognition

List of references (Incomplete)

Those especially recommended as textbooks are marked with ★. The contents related to this class are listed as [ ].

- On Phonetics
  - Ken Machida Editor, Hajime Inotsuka, Emiko Inotsuka, 'The mechanisms of Japanese phonetics ' (series-Exploring the mechanisms of Japanese language) KENKYU-SYA, 2003, ISBN4-327-38302-3, 2000 YEN
  - Haruo Tsubosono ‘The Speech in Japanese' (Introduction to modern linguistics 2), IWANAMI-SHIN-SYO, 2001, ISBN4-00-006692-7 C3380, 3400 YEN
- On Speech Information Processing
  - Hiro Moriya ‘Speech coding', Institute of Electronics Information and Communication Engineeres, 1998, ISBN4-88552-156-4, 3000 YEN ★★
  - Akio Ando, 'Real Time speech recognition', Society of Electron Information Communication, 2003, ISBN4-88552-195-5, 3600 YEN ★★
  - Shikano, Ito, Kawahara, Takeda, Yamamoto, 'speech recognition systems' (Information Processing Society of Japan IT text), OUM-SHA, 2001, ISBN4-274-13228-5, 3500 YEN ★
  - Shikano, Nakamura, Ise, ‘Speech / Acoustic Information Digital Signal Processing’, SHO-KEI-DO, 1997 ★
  - Masao Kasuga, Tetsuo Funada, Shinji Hayashi, Kazuya Takeda, ‘Speech information processing' (The Institute of Image Information and Television Engineers Essential Technology series 1), CORONA-SHA, 2001,ISBN4-339-01261-0, C3355, 3500 YEN

- On Signal Processing
  - Takashi Soeda, Takayoshi Nakamizo, Shigeru Omatsu, ‘Fundamentals and Applications of Signal processing', (Fundamentals in Science and Engineering Series) NISSHIN-SYUPPAN, 1979, ISBN4-8173-00106-6 C0041, 3000 YEN
  - Hiroshi Kanai, 'Spectrum analysis of Sound and Vibration', (edited by Acoustical Society of Japan, Sound Technology Series 5), CORONA-SHA, 1999, ISBN4-339-01105-3 C3355, 5000 YEN〔Distances, z-transform, DFT, LPC, PARCOR〕
  - Masafumi Hagiwara, 'Digital signal processing', (Electronics, Information and Communication Engineering Series), MORIKITA-SYUPPAN, 2001, ISBN4-627-70131-4 C3355, 2000 YEN
  - Hijiri Imai, 'Signal Processing Technology of Signal/System Theory and Processing Technology-', (edited by the Institute of Image Information and Television Engineers, Manual of the Television Society Series 8), CORONA-SHA, 1997, ISBN4-339-01058 C3355, 2800 YEN
  - Nozomi Hamada,'Easy to understand signal processing', (Semesta training series), OUM-SHA, 2000, ISBN4-274-12990-X C3055, 2400 YEN
  - Shigetoku Kaneshiro, HIroshi Ochi, ‘Learning Digital Signal Processing through examples', CORONA-SHA, 1997, ISBN4-339-00678-5 C3055, 2400 YEN
  - Shigeo Tsujii, Hajime Kubota, 'Easy to learn Digital Signal Processing', OUM-SHA, 1993, ISBN4-274-12939-X C3055, 2500 YEN
- On Statistical Mathematics
  - Michiko Watanabe, Kazunori Yamaguchi Editors, 'EM algorithm and problems of incomplete data', TAGA-SYUPPAN, 2000, ISBN4-8115-5701-8 C1033, 6600 YEN

[FYI]Contents of "Special Course in Signal Processing" in the Graduate School

- Review of the undergraduate class
  - Cepstrum→MFCC
  - HMM
- Spoken word recognition
- *n*-gram language model
- Large vocabulary speech recognition, Continuous speech recognition
  - Basic algorithms
  - Training: EM-algorithm
  - Search: A* search
- Phonetical models
  - GMM
  - Tying［Young］
  - Context-dependent phoneme model［K-F Lee］
- Language models
  - Network grammar
  - Context-Free Grammar (CFG)
  - n-gram statistical grammar
  - Latent Semantic Analysis［Belegarda］
- Search technique
  - *N*-best algorithm (tree-trellis)［Lee & Soong］
  - Stack decoder
- Channel adaptation
  - CMN［Acero & Stern］
- Speaker adaptation
  - VFS［Sagayama］
  - MAP［Lee］
  - MLLR［Legetter］
  - EigenVoice［Kuhn］
- Noise adaptation
  - SS［Noll］
  - Varga-Moore Method［Varga］
  - PMC/NOVO［Gales］
  - JA［Sagayama］
- Microphone array
  - Sound models training methods
- ML, MAP, MMI, MCE
- Speech synthesis, speech conversation
  - Speech synthesis, HMM synthesis, prosody models
  - Anthropomorphic spoken dialog agent
  - Future issues of speech recognition
  - Spontaneous speech
  - Automatic speech translation

Department of Mathematical Engineering and Information Physics and speech research

- A significant number of speech researchers came from the department of Mathematical Engineering and Information Physics

- Engineering of physical information (measuring, mathematics, etc.)

  Cf. Today's 'Information Engineering': Discrete Information, Symbolic Information are the main parts.

- Students of Department of Mathematical Engineering and Information Physics are 'all-rounders'

Why speech research? – My(Sagayama's) case

- During middle high school
  - English: fascinating phoneticl symbols [æ, ʌ, ɔ, θ, ð, ʃ, ə, ɚ, ɛ, …]
  - Japanese: oral grammar
- During high school
  - Chorus: Latin (r, ɾ, ɹ, ʍ, ʀ, ʁ, …)
  - Fascinating phonetical symbols of Professor Higgins in the musical, 'My fair lady'
- In university
  - (Phonetics)
  - Undergraduate project: pattern recognition
  - Graduate school project: acoustic measurements (M-sequence modulated correlation method), signal processing
- At the NTT Electrical Communication Research Laboratories (later renamed NTT Human Interface Research Laboratories)
  - Mathematical techniques
  - Worked with active researchers (Shuzo Saito, Shinichiro Hashimoto, Fumitada Iikura, Tadanori Kohda, Hirokazu Sato, Sadaoki Furui, Ryohei Nakatsu, Nobuhiko Kitawaki, Kiyohiro Shikano, Yoichi Tohkura)
- At ATR Interpreting Telephony Laboratory (as Head, Speech Information Processing Laboratory)
  - Leading edge frontline of speech recognition and synthesis technology
- 'In the beginning was the Word, and the Word was with God, and the Word was God.' [John, The Holy Bible]
  - In fact, 'word' must have meant 'speech' when there were no letters. (that means speech research is …)