Spoken Languages Systems: Introduction, Issues and Current Research

Roberto Togneri
CIIPS (SIP Lab)
EE&C Eng, UWA
Overview

- Schematic of Spoken Language System (SLS)
- Introduction to SLS
  - Feature extraction
  - Acoustic Modelling
  - Language Modelling and Understanding
- Issue of Robustness
  - Hidden Dynamic Model
  - Robust Signal Processing
  - Current status
- Concluding Remarks
Schematic of SLS

- Training/Estimation (EM algorithm)

- Feature Extraction

- Baum-Welch

- Acoustic Model

E-step

M-step
Schematic of SLS

- Testing/Decoding (Viterbi algorithm)
Schematic of SLS

- Language Understanding and Processing

ASR Module

NLP Module

TTS Module

Information Retrieval

Response Display

N-best list

semantic frames
Feature Extraction

What is feature extraction?
- Input is acoustic signal, $s(t)$
- Output is a vector of features for each analysis time frame
- Extract features relevant to what is being said
- Ignore features related to speaker and environment

Engineering considerations
- Robustness to noise (environment)

Acoustic-Phonetic considerations
- Unit of speech (i.e. phoneme)
- Co-articulation effects
Acoustic Modelling

- What is acoustic modelling?
  - Input are features
  - Output are models of speech
    - phone, tri-phone, syllable, word

- Engineering considerations
  - Identification of model parameters
  - Uniqueness of model

- Mathematical considerations
  - Statistical estimation and evaluation

- Acoustic-phonetic considerations
  - co-articulatory (i.e. tri-phone models)
Language Modelling

- What is language modelling?
  - Recognition requires a search through all possible word sequence combinations
  - Language model restricts search to most plausible word sequences

- Language considerations
  - For given sequence of words what are the most likely words to follow?
  - For given task or context what are the most likely sequences? (grammar)

- Mathematical considerations
  - Statistical language models (N-gram)
Language Understanding

- What is Language understanding?
  - Unknown utterance has been transcribed into a sequence of words, with possible insertion and deletion errors
  - NLP extracts the “meaning” of a word sequence
    - robust NLU to cope with errors
    - initiates action or response
  - Dialogue management
    - Produces follow-up queries to extract the information needed for the specified task
    - iterations of TTS and ASR
State-of-the-art HMM

- Models time evolution of speech
  - Each state corresponds to quasi-stationary sound (phoneme!)
  - Stochastic nature of transitions model “slow” and “fast” speech dynamics
  - Stochastic nature of state distributions model “noisiness” of acoustic features
Limitations of HMM

- Modelling Limitations
  - Assumes acoustic features are independently generated
    - only true for static voiced sounds
    - continuous speech is never static!
  - State duration and transition likelihoods do not follow that of a Markov model
  - Poor modelling of plosive sounds
  - Data-driven modelling approach
    - Lack of structure implies more models, many parameters and a lot of reliable data to be available
    - CD phones (tri-phones) needed to model effects of co-articulation
Hidden Dynamic Model

- **HDM Fundamental Idea**
  - hidden dynamic based on speech data generation
    - hidden dynamic can represent articulatory or vocal tract resonance (VTR) dynamics
  - context-dependence inherent in the model structure
    - continuity condition of hidden dynamic

- **VTR dynamics**
Hidden Dynamic Model

- Non-linear switched target-directed hidden dynamic state-space model
  \[ Z_{k+1} = \Phi_j Z_k + (I - \Phi_j) T_j + w_k \]
  \[ O_k = h_j(Z_k) + v_k \]

- \( T \) describes VTR of phone
  - formants in the case of voiced sounds
- \( \Phi \) describes “speaking style” of phone
- \( Z \) is continuous from phone \( j \) to \( j+1 \)
  - inherent context-dependence modelling
- \( h(Z) \) represents a mapping from the VTR dynamic to the observations
Hidden Dynamic Model

What is needed?

- From engineering and mathematics
  - more efficient and reliable estimation and decoding algorithms from the statistics, communications and control engineering literature
  - a more constrained mapping function between hidden dynamics and observable acoustic features, MLP is too general!

- From speech physiology
  - more data on time-constant and target parameter values, different dynamics and parameters

- From acoustic-phonetics
  - more information on time-constant and target parameter values, different dynamics and parameters
Robust Signal Processing

- Speech Signals are subject to noise
  - Additive noise
  - Channel noise (linear filtering)
  - Reverberant noise
  - Stressed Speech

- State-of-the-art feature processing
  - Mel-Frequency Cepstral Features with CMN
    - CMN eliminates channel noise
    - CMN features are decorrelated, and the speaker normalised (speaker pitch can be “liftered” out)
    - CMN features affected non-linearly by added noise
Robust Signal Processing

- State-of-the-art feature processing
  - Array processing (also BSS)
    - Multiple microphones permit speaker localisation
    - Effective against reverberant noise
  - Spectral Subtraction
    - Subtract noise spectrum in the spectral domain
    - Effective against additive noise
- Model adaptation
  - HMM models are adapted to the noisy environment
  - Requires adaptation data
  - Need new adaptation for each different environment
Robust Language Understanding

- Robustness Problems
  - Spontaneous Speech
    - Disfluencies: “uh” “um” “err”
    - Restart: “I want .. err .. I would like”
    - Auto-correction: “I would like to arrive depart Boston”
  - Grammatically Incorrect
    - “departing on eight thirty am”
  - Word errors from ASR module

- Solutions
  - Probabilistic grammars
  - Noise heuristics
Conclusions

- Hot research areas in SLS
  - Robustness to noise
    - Reverberant noise
    - Additive non-stationary noise
  - Advanced Acoustic Modelling
    - Hidden Dynamic Model
    - Need to handle Spontaneous Speech
  - Robust Language Understanding

- Current projects
  - Robust feature extraction
  - Hidden Dynamic Model
  - Robust language modelling / understanding