Overview

• Characteristics of the Speech Signal
  – A continuous-valued time series generated by encoding various of excitation with a complex time-varying non-linear filter.
  – various kinds of energy excited by

• Multi-Class Extensions
  – combining binary SVMs
  – multi-class SVMs

• Structured SVMs for Continuous Speech Recognition
  – joint feature spaces for structured modelling
  – large margin training
  – relationship with other models
  – lattice based implementation
Characteristics of the Speech Signal

• A continuous-valued time series generated by encoding various of excitation with a complex time-varying non-linear filter.
  – Continuous-valued: impact on our choice of models and need to be careful with the numerical computation.
  – Time series: the model need to be able to represent this, and the training and decoding efficiencies are often of concern.
  – Speech signals are presented in the form of rapidly-varying functions.

• Speech signals produced by humans are often pre-processed with signal processing methods and used as the input features to the automatic speech recognition (ASR) system.
  – ASR need to handle the variability of humans: coarticulation, time-varying (mood, aging, ...), gender, accent, and etc.
  – ASR need to face difficulties existed in the other signal processing methods: channel variations, noise, ....
Resources Available for Building ASR

• Phonetic knowledge characterizing how phones are produced with articulator movements.
  – Some rules need to be verified across a large amount of speakers.
  – State-of-the-art ASR often adopts statistic models trained with a large amount of speech data (e.g., 3000 hours – 1.08G samples).

• Lexical and syntax knowledge is available for a given language and can aid speech recognition.
  – Our-of-vocabulary words.
  – Ill-formed sentences.
Some Basis of Stochastic ASR

- Continuous speech signals are sampled to discrete waveforms, then compressed to a sequence of individual speech frames according to the short-time stationary property (10~30ms/sec), assuming the vocal tract is time-invariant.

- Source-filter model based on maximum a posteriori criterion,

\[ \hat{w} = \arg \max_w P(w|O) \propto \arg \max_w P(O|w)P(w). \]

- \( O \) refers to the input speech frame sequence, \( w \) refers to the word sequence.
- \( P(w) \) and \( P(O|w) \) are called the language model and the acoustic model.
- \( \arg \max_w \) is to decode for the most likely hypothesis.

- Hidden Markov Models (HMMs) are most commonly used under the framework.
The sound of a phonetic unit can often be divided into several states, denoted as $s$, according to its production procedure. Assume $s$ is 1st-order Markovian,

$$P(s) = \prod_{t=1}^{T} P(q_t = s_t, q_{t-1} = s_{t-1}).$$

It is sensible to regard the phone as produced by another process associated to $s$. Let us assume the process only depends on the current state, i.e.,

$$P(O|s) = \prod_{t=1}^{T} P(o_t|s) = \prod_{t=1}^{T} P(o_t|q_t = s_t).$$
Now we have a HMM, denoted it as $\lambda$,

$$P(O|\lambda) = \sum_s P(O|s, \lambda)P(s|\lambda)$$

In ASR, we usually use constant transition probabilities between different states, denoted as $P(q_t = s_t|q_{t-1} = s_{t-1}) = a_{t-1,t}$.

Modern ASR uses continuous density to model the observation probabilities. Assuming the frames belong to a certain state are i.i.d, Gaussian mixture models are commonly used to approach any continuous density associated with that state by any precision, i.e.,

$$b_j(o_t) = P(o_t|q_t = j) = \sum_{m=1}^{M} c_{jm}N(o_t, \mu_{jm}, \Sigma_{jm}).$$
HMM Acoustic Models & Decoding

- A set of acoustic models contains HMMs relevant to every phone (syllable, word, and *etc.*) of the target language.

- Modern ASRs use a tuple of concatenated phones rather than a single phone to build an HMM, to capture coarticulation changes inter/intra words (e.g., triphone: ‘IY’ ‘T’ ‘CH’ ‘IY’ ‘Z’ → ‘sil’+‘IY’-‘T’ ‘IY’+‘T’-‘CH’ . . .)
  - Relevant states to triphone HMMs with the same central unit are often clustered to avoid data sparseness and reduce system complexity.
(Deep) Neural Networks in ASR

• To our knowledge, DNN applications in ASR (in addition to LM) include 3 aspects:
  – Acoustic models: use the pseudo posteriors from DNN to obtain the observation probabilities.
  – Tandem feature detectors: to extract discriminative neural net features and use them together with the original observations.
  – Speech attribute detectors: use DNNs to extract a set of asynchronous speech attributes.

• The DNN most commonly used in ASR is deep feedforward NNs (expect for LM, where people also use deep recurrent NNs).

• The training approaches in use include:
  – Layer-wised generative pre-training (RBM and etc.)
  – Layer-wised discriminative pre-training.
  – Normalized random initialization.
  – 2nd-order optimization.
DNN-HMM Acoustic Models

- A DNN with phone or tied-state targets (√) is fitted into HMM acoustic models by converting the pseudo posteriors into the observation probabilities,

\[
\ln P(o_t|s_t) = \ln P(s_t|o_t) - \ln P(s_t) + C,
\]

where \( C \) is a negative constant, \( C \propto \ln P(o_t) \).

- Comparing DNN-HMM acoustic models to GMM-HMM acoustic models,
  - GMMs are trained generatively (needs an additional pass of discriminative training to be discriminatively), individually, and sequentially.
  - A DNN is trained discriminatively and globally on frame-level (also can be trained on sequence level by back-propagating the statistics generated and collected using sequential criterion).
  - A DNN can take the observations of several concatenated frames as the input directly, utilizing the context information.
Tandem Feature Detectors

• The way of using tandem features:
  – Extract neural net features.
  – Combine the neural net features with the original input observations.
  – De-correlate and reduce the dimensions of the tandem features.
  – Use tandem features rather than the original observations as the input to the diagonal GMM-HMM acoustic models.

• Different kinds of DNN features:
  – DNN output posteriors: phone posteriors and tied-state posteriors.
  – Bottleneck DNN: build a DNN (either phone or tied-state targets) with a bottlenecked hidden layer; use the linear output of the bottleneck layer as the DNN features.

• GMM-HMM systems with DNN (tied-state posteriors) bottlenecked tandem features are reported to have comparable performance to DNN-HMM systems.
Speech Attribute Detectors

- Some researchers claim the linear-chain structure of HMMs is not suitable to cover speech variations, and it may ignore some useful knowledge. Therefore proposed to use detection-based system.
  - Extract and utilize various of features from the speech signals based on prior knowledge from linguistics, signal processing, neuroscience, ...  
  - To use more complex model and system structure.
  - The accuracy of detectors was a key factor impact on the performance.

- Recent studies utilized DNN to detect articulation derived speech attributes, and got good results.