Automatic Speech Recognition

Claude Barras

barras@vocapia.com

Master AI - Nov. 2020

VOCAPIA research

Plan

VOCAPIA research

Plan

VOCAPIA research

Introduction

VOCAPIA research

- Human-machine spoken communication - why?
  - Spontaneous, faster than keyboard
  - Hand-free human-machine communication
- What can we extract from speech?
  - Language identification
  - Speaker recognition
  - Speech transcription
  - Information on affective state and health
- Automatic speech recognition (ASR) is complex
  - No space between words (coarticulation)
  - Variable temporal flow
  - Inter- and intra-speaker variability
  - Homophones...

Short history of ASR

VOCAPIA research

- 1960s
  - Dawn of AI: experimental rule-based systems
- 1970s
  - Pattern recognition: isolated word recognition with DTW
- 1980s
  - Statistical approaches: continuous speech transcription with HMM
- 1990s
  - International evaluation campaigns, corpora collection
- 2000s
  - Rise of DNN
- 2010s
  - Generalization of consumer voice assistants by GAFAM

Plan

VOCAPIA research

Isolated word recognition

VOCAPIA research

- Pattern recognition framework
  - R_w reference for each word w
  - X unknown observation
  - D a distance between acoustic segments
  - choose $\tilde{w} = \arg\min_w D(X, R_w)$
- Acoustic segments
  - variable-length *sequences* of acoustic samples
  - waveform amplitudes too noisy for reliable comparisions
  - replace with a sequence of acoustic vectors (100 vec/sec)
  - 10-15 dim vectors from spectral analysis of 10-30ms frames
- Inter-segments distance $D(X, Y)$ with $X = \{x_i\}_{i=1\ldots N}$, $Y = \{y_j\}_{j=1\ldots M}$
  - Rely on local inter-vector distance $d(x_i, y_j)$
  - Due to non-linearity, need to find the best alignment
  - Choose best path P for $D(X, Y) = \min_P \sum_{(i,j)\in P} d(x_i, y_j)$
  - Combinatorial explosion of problem with segment sizes!

Dynamic Time Warping

VOCAPIA research

- Dynamic programming
  - Shortest path between nodes in a weighted graph
  - Bellman optimality principle: subpath of optimal path is optimal
  - Allow optimal recursive resolution (eg. Dijkstra's shortest path)
- Dynamic Time Warping (DTW)
  - Express the elactic distance D in terms of shortest path in a graph
  - Complexity linear with segment lengths N and M
- Simple algorithm
  - Initialization: $D_{0,0} = 0$, $D_{i,0} = D_{0,j} = \infty$
  - Recurrence: $D_{i,j} = \min_{(k,l) \in \text{prec}(i,j)} D_{k,l} + d(x_i, y_j)$ with $\text{prec}(i,j)$ a small set of preceding indices
  - Final step: $D(X, Y) = D_{N,M}$
- Application
  - Keyword-based voice command

Plan

VOCAPIA research

Statistical paradigm

VOCAPIA research

- Problem
  - Let $X = \{x_1 \dots x_T\} = x_1^T$ the signal
  - Find the best matching possible sentence w^*

$$w^* = \arg\max P(w_1^n \mid x_1^T)$$

  - Generative view (source through noisy channel) using Bayes rule

$$w^* = \arg\max P(x_1^T \mid w_1^n) P(w_1^n)$$

- Three sub-problems
  - Acoustic model $P(x_1^T \mid w_1^n)$
  - Linguistic model $P(w_1^n)$
  - Search algorithms for argmax

Continuous speech recognition

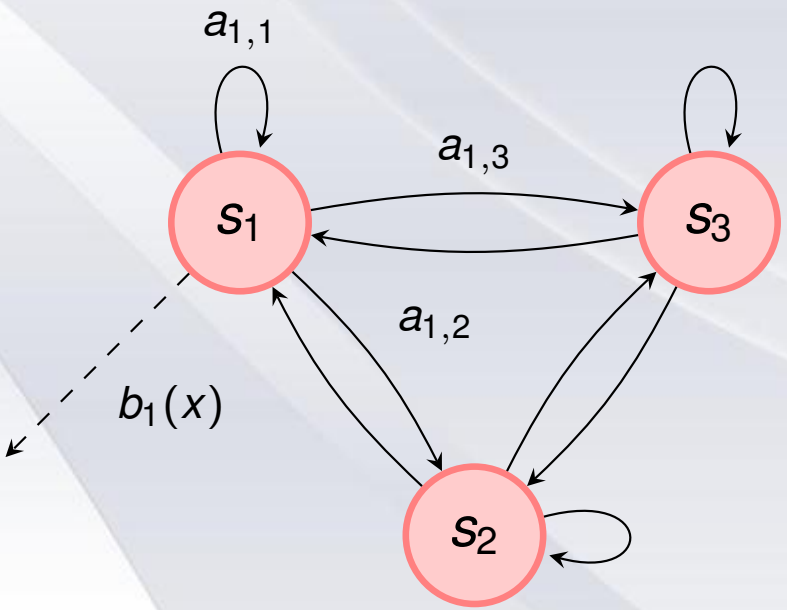
VOCAPIA research

- Large vocabulary continuous speech recognition (LVCSR)
  - Impossible to train a model for each word
  - Instead, select short acoustic units, typically phones in context
  - Select a limited vocabulary V
  - Map each word to its phonetic pronounciation
  - Deterministic or probabilistic model of V^*
  - Heuristic search into $H \subset V^*$
- Resources and training data
  - x 100h of audio with precise and synchronized manual transcription
  - texts x100M up to 1G words
  - dictionary with pronounciation(s)

Hidden Markov models (HMM)

VOCAPIA research

- Hidden Markov model (HMM)
  - a stochastic state machine with random drawing of acoustic vectors
- defined by
  - S a set of states s_i, with process in state q_t at time t
  - A the inter-state transition matrix
    $a_{i,j} = P(q_t = s_j | q_{t-1} = s_i) = P(s_j | s_i)$
  - B the observation distribution into acoustic space X
    $b_i(x) = P(x | q_t = s_i) = P(x | s_i)$



Hidden Markov models (HMM)

VOCAPIA research

- Decoding, given a HMM model λ
  - Find the most likely sequence of hidden states for X?

$$q^*_{1,T} = \arg\max q_{1,T} P(q_{1,T}|x_{1,T}, \lambda)$$

  - Dynamic programming resolution with Viterbi algorithm
- Compute $P(X|\lambda)$?
  - Forward-backward algorithm (variant of Viterbi)
  - Allow to select the best-matching model
- Training of model
  - Given a set of sequences $\{X_k\}$, optimize the model parameters

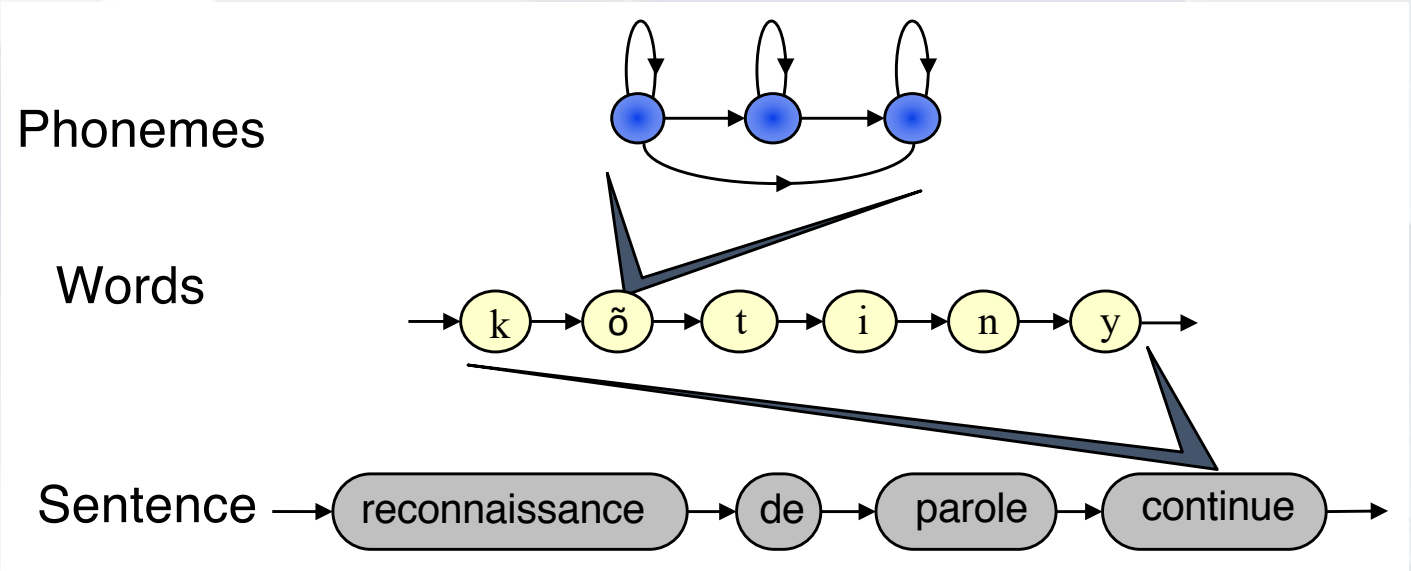
$$\lambda^* = \arg\max_{\lambda} \sum_k P(X_k|\lambda)$$

  - Expectation-Maximization Baum-Welch iterative algorithm
  - Improve the model at each iteration (may get stuck in local extremum)

Acoustic modeling

VOCAPIA research

- Compute $P(X|W)$ with generative HMM models
  - Each phone has a left-to-right topology for temporal causality
  - Few states, each corresponding to a short acoustic segment
  - Generation of acoustic vectors with Gaussian mixture models (GMM)
  - Model of sentence by hierarchically embedding words and phonemes



Acoustic modeling

VOCAPIA research

- Triphones
  - 40 phones => 64.000 triphones in left/right context: impossible to model all properly
  - Tying of states (sharing parameters) for similar contexts using linguistic rules
  - Model size: 10.000 shared states x 16 Gaussians x 2 (mean + diagonal covariance) x 39 (dimension of acoustic vectors) $>$ 10M param
- Adaptation of acoustic models
  - Multi-speaker acoustic models trained on a large database
  - Need to adapt them to speaker and/or acoustic conditions
  - Unsupervised adaptation (without transcripts) more convenient
  - Various approaches: VTLN (vocal tract length normalization), MAP (Maximum a posteriori), MLLR (Maximum likelihood linera regression)...

Linguistic modeling (LM)

VOCAPIA research

- Formal grammars
  - Rules covering all possible sentences
  - Suited to artificial languages or very constrained domains
- Probablilistic grammars
  - n-gram model
  - Data-based, simple but relying on large corpora
- n-gram
  - Formally, $P(W) = \prod_{i=1}^{n} P(w_i|w_1 \ldots w_{i-1})$
  - Untractable, so limit to short-term history
  - 1 word for bigram: $P(W) \simeq \prod_{i=1}^{n} P(w_i|w_{i-1})$
  - 2 words for trigram: $P(W) \simeq \prod_{i=1}^{n} P(w_i|w_{i-1}, w_{i-2})$
  - Maximum likelihood estimation through counts of words sequences
  - LM quality measured through perplexity (related to model entropy)
  - Limitations related to Zipf law (few frequent words, frequent rare words)
  - LM needs to predict unseen word sequences
  - Workaround is model smoothing through interpolation or back-off

HMM decoding

VOCAPIA research

- Viterbi algorithm
  - Embed acoustic HMM and linguistic n-gram
  - Potentially huge search graph
  - Rely on dynamic programming, but not enough
  - Heuristics needed to prune the graph: eg. beam search (discard hypothesis too far from the current best one)
  - Careful balance between speed and accuracy (avoid too much pruning)
- Multi-step decoding
  - Produce n-best hypothesis rather than 1-best
  - First output a word graph (lattice) - allow a fast rescoring with better models
  - Compress the graph accross time in a consensus network

Plan

VOCAPIA research

Neural networks for ASR

VOCAPIA research

- Lot of research since the 1990s with multilayer perceptron (MLP) and recurrent networks
  - Bourlard, Robinson, Bengio, Gallinari, Waibel...
- No significant gain compared to "standard" probabilistic systems (GMM/HMM) for years
- Integration into SOTA systems for linguistic (>2002) and acoustic (>2006)
- Decoding mostly relying on dynamic programming (Viterbi or similar) with development of CTC (connectionist temporal classification) (Graves, 2006).

Neural networks for ASR

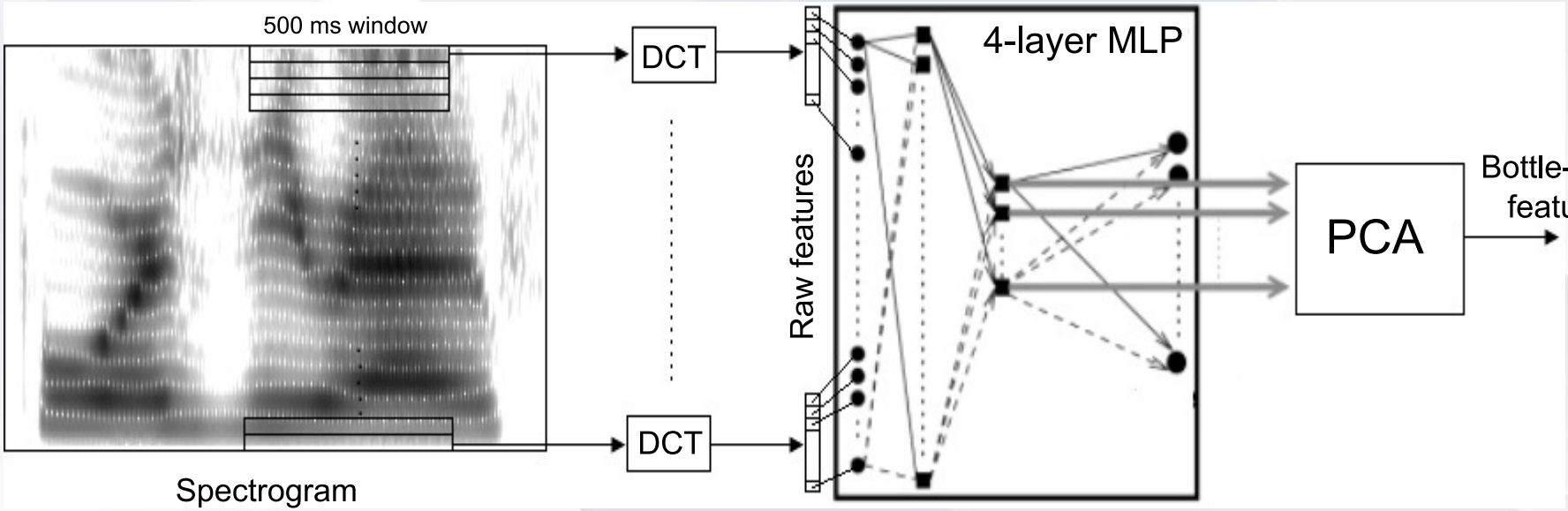
VOCAPIA research

- Neural Linguistic models
  - Proposed by Y. Bengio in 2001
  - Integration into LIMSI since 2002 combined with n-gram
  - Significatif gain (5% relative gain in accuracy)
- Neural acoustic models
  - Hybrid systems MLP/HMM
  - Estimation of phonetic posterior probabilities by a MLP from the acoustic features
  - MLP output replaces the GMM in a HMM state
  - Another solution: features output from the MLP are combined with standard acoustic features (MFCC)
  - The rest of the system is usually a standard HMM

Hybrid acoustic system - an example

VOCAPIA research

- TRAP-DCT (Grezl & Fousek, 2008)
- Input : 19 bands x 25 coefficients
- 3rd layer (bottleneck) : 39 coefficients + PCA (decorrelation)
- 4th layer (output) : phonetic states probabilities



Neural networks for ASR

VOCAPIA research

- Important development of neural approaches with deep networks (DNN)
  - convolutive networks (CNN) connected to a spectral bank filter
  - recurrent network (Bi-LSTM = bi-directionnal long-short-term memory networks) for audio stream
  - applied to consumer voice assistants, keyword detection, speaker recognition, speech synthesis
  - efficient software toolkits for generic DNN (TensorFlow, PyTorch...) or dedicated to speech recognition (Kaldi).
- Examples
  - `https://machinelearning.apple.com/2017/10/01/hey-siri.html`
- Recent trend towards end-to-end (E2E) ASR
  - New DNN architectures: encoder-decoder with attention (Transformer)
  - `http://iscslp2018.org/images/T4_Towards%20end-to-end%20speech%20recognition.pdf`