

Automatic Speech Recognition: Introduction, Current Trends and Open Problems

Une brève histoire de la reconnaissance de la parole
(mais qui commencerait seulement dans les année 90 ...)

Laurent Besacier



- 1 The speech signal
- 2 1990-2015: Bayes, HMMs, GMMs
- 3 2015-: neural nets
- 4 Is ASR a solved problem ?

- 1 The speech signal
- 2 1990-2015: Bayes, HMMs, GMMs
- 3 2015-: neural nets
- 4 Is ASR a solved problem ?

Speech facts

- Speech generally conveys a (linguistic) message (that can be reduced to a transcript)
- But not only (paralinguistics: speaker identity, speaker mood, speaker health condition, speaker accent, etc.)
- Variability at all levels (intra speaker, inter speaker, microphone, phone line, room acoustics, style)
- Speech is a continuous signal (no explicit word boundaries)
- May be decomposed into elementary units of sound (phonemes) that distinguish one word from another in a particular language (minimal pairs)
 - *kill* vs *kiss* - *pat* vs *bat*
 - phoneme set is language dependent
 - acoustic realization of the phoneme is dependent of its left and right neighbors (co-articulation)

(Main) Speech tasks

- Speech compression (solved)
- Speaker recognition (strong progresses over the last 10 years but still poor compared to other biometric modalities like fingerprint and iris)
- Text-to-speech synthesis (can still gain in naturalness but new progresses with DL: Wavenet, Tacotron2, VoiceLoop)
- **Speech-to-text (this talk)**
- Speech paralinguistics (early days): detection of gender, age, deception, sincerity, nativeness, emotion, sleepiness, cognitive disorders, (drug or alcohol) intoxication, pathologies, etc.
- Main speech conference: *Interspeech* (core A, every year)

Speech-to-text

- Automatic Speech Recognition (ASR)
- Ideally we want to have a system that deals with: spontaneous speech, multi-speakers, unlimited output vocabulary, any acoustic condition
- But performances differ greatly for different contexts (read vs spontaneous speech ; small vs large vocabulary ; quiet vs noisy)

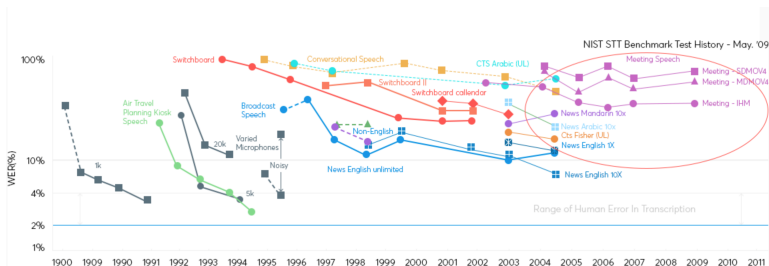


Figure: NIST ASR benchmark tests history (< 2015)

ASR as a downstream task

- ASR for spoken language processing (speech understanding, speech translation, speech summarization, etc.)
- Not just a problem of noisy transcripts
- No sentence boundaries, punctuation, case
- Disfluencies in spontaneous speech: false starts, fillers, repaired utterances
 - btw, should we keep them or remove them ?
 - some speech tasks are ill defined (ex: speech translation)
- Time to work on end-to-end approaches from speech ?

Speech representations

- Handcrafted feature vectors
 - standard extraction on sliding windows of 20-30ms at a frame rate of 10ms
 - filterbanks (signal energy in different frequency bands)
 - cepstral coefficients (inverse Fourier transform of the logarithm of the estimated spectrum of a signal)
 - linear predictive coding (a sample is predicted as a weighted sum of preceding samples and weights are used as features)
 - prosodic features (pitch, energy)
- Raw waveform (> 2015)
 - bypass handcrafted preprocessing
 - preprocessing become part of the acoustic modeling and training
 - introducing convolutional layers in the first stages of the NN pipeline

Speech representations

- Spectrograms (< 1990 and $> 2015!$)
 - time-frequency representation that is actually similar to sequence of filterbanks ...
 - ... but processed as an image

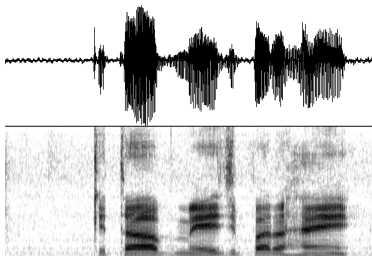


Figure: Speech signal (top) and spectrogram (bottom)

Progresses over the years

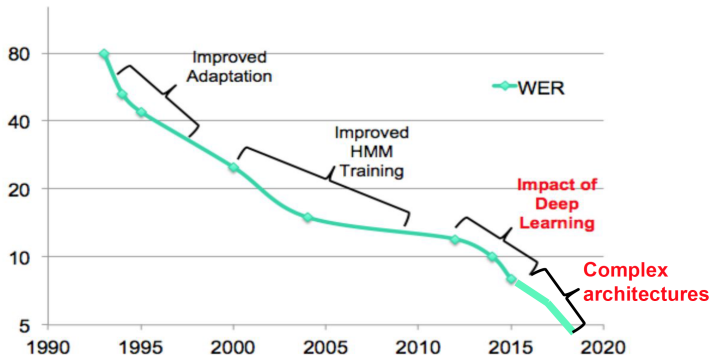


Figure: ASR Performance¹ on English Conversational Telephony (Switchboard)

¹Image from Bhuvana Ramabhadran's presentation at Interspeech 2018

- 1 The speech signal
- 2 1990-2015: Bayes, HMMs, GMMs
- 3 2015-: neural nets
- 4 Is ASR a solved problem ?

Fundamental equation

x : observation (signal or features)

w : a word sequence

$$w^* = \operatorname{argmax}_w p(w/x) = \operatorname{argmax}_w p(x/w) \cdot p(w) \quad (1)$$

$p(x/w)$: acoustic model

$p(w)$: language model

Lexicons

- For acoustic modelling in large vocabulary speech recognition, we model phones instead of full words
- A pronunciation lexicon gives the decomposition of words into phonemes
- Adding a new word to the output vocabulary does not require retraining of the acoustic models
 - just add an entry to the pronunciation lexicon
 - *cat* /k a t/
- Hierarchical modelling of speech (signal/phones/words/utterance)

Hierarchical modelling of speech

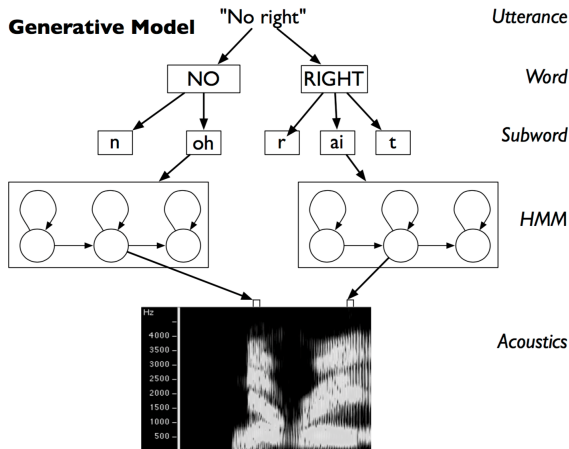


Figure: From speech to utterances²

²Image from Steve Renals's lecture on ASR

ASR overview

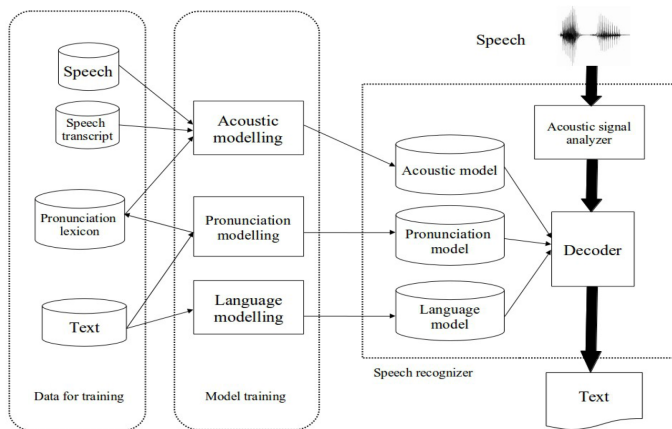


Figure: ASR Overview

Acoustic modeling: HMM/GMM

- Complex sequential patterns of speech decomposed into piecewise stationary segments
- Sequential structure of the data described by a sequence of states
 - HMM (Hidden Markov Models) transitions
- Local characteristics of the data described by a distribution associated to each state
 - GMM (Gaussian Mixture Models) observations (outputs)

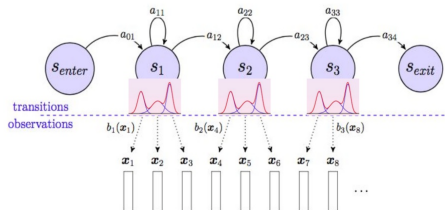


Figure: HMM/GMM approach

Language models: from N-grams to RNNs

For a sequence of T words $W = w_1, w_2, \dots, w_T$

$$P(W) = \prod_{k=1}^T P(w_k | w_1, w_2, \dots, w_{k-1}) \quad (2)$$

$$P(W) = \prod_{k=1}^T P(w_k | h) \quad (3)$$

n-gram LM: $h = w_{k-n+1}, w_{k-n+2}, \dots, w_{k-1}$

recurrent neural network LM: $h = rnn_state(E(w_1), E(w_2), \dots, E(w_{k-1}))$

- 1 The speech signal
- 2 1990-2015: Bayes, HMMs, GMMs
- 3 2015-: neural nets**
- 4 Is ASR a solved problem ?

NNs in the 90s and 00s

- Introduced in the 80s and 90s to speech recognition, but extremely slow and poor in performance compared to the state-of-the-art HMM/GMM
- Several papers published by ICSI, CMU, IDIAP several decades ago!
- Pros: no assumption about a specific data distribution
- Cons: slow and do not scale to large tasks

Deep learning breakthrough

Like in vision, due to

- More data
 - ex: (2015) Librispeech (en) 1.000h (Panayotov et al., 2015)
 - ex: (2016) Baidu Deep Speech 2 (en) 12.000h (Amodei et al., 2016)
 - ex: (2017) Google Home (en) 18.000h (from a Google presentation)
 - ex: (2018) Google wav2words (en) >100.000h?³ (informal discussion)
- Computation (ex: GPU)
- Better optimization algorithms and training objectives
- ASR Toolkits (ex: Kaldi (Povey et al., 2011) and DL frameworks (Tensorflow and the like))

³>11 years of speech !

End-to-end ASR (get rid of HMMs)

Two approaches for end-to-end ASR

- Connectionist Temporal Classification (CTC)
 - Solves the problem of unaligned input and output sequences by marginalizing the conditional likelihood of the output sequence given the input over all possible alignments
- Attention Modeling
 - Simultaneously optimize alignment and grapheme (or word) decoding using attention weights (linear combination of hidden states) to influence the generated output

CTC

- Graves et al. (2006) introduced the CTC loss function
- Defined over a label sequence z (of length M)
- blank* or $_$ symbol allows M -length target sequence to be mapped to a T -length sequence x
- z can be represented by a set of all possible CTC paths (sequence of labels, at frame level) that are mapped to z
 - ex: $M=2$ ($z = hi$) and $T=3$ (3 frames): possible sequences are 'hhi', 'hii', '_hi', 'h_i', 'hi_'
- Probability $p(z/x)$ evaluated as sum of probabilities over all possible CTC paths (using Forward-Backward)
- Generate frame posteriors at decoding time

Per-frame argmax:

```

yy oe      tt      a
rr e      hh      ll i      tt      aa      tt      iio n
cc rrr u      ll ss      nddd      i n
th e      nn      hhh a      nng
l o o g g ii nng
b rr ii      ck s      p ll a      smt      eerr
a nnd      b ll uu      ee pp      r i      mss      f oo rrr tt y
t www oo      nn ew      b e      t i n
e pp      aa rr tt      mm ee      nntss

```

After collapsing:

yet a rehabilitation cru is onhand in the building loogging bricks plaster and bluepains four forty two new betin aptments

Attention modeling

- Architecture similar to neural machine translation
- Speech encoder based on CNNs or pyramidal LSTMs ?

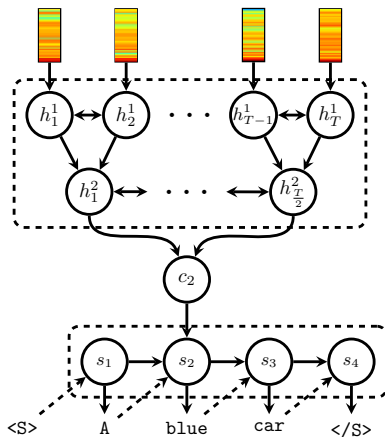


Image from Alexandre Berard's thesis

- 1 The speech signal
- 2 1990-2015: Bayes, HMMs, GMMs
- 3 2015-: neural nets
- 4 Is ASR a solved problem ?

On par with human transcription ?

| Read Speech | | | |
|------------------------|-------|-------|-------|
| Test set | DS1 | DS2 | Human |
| WSJ eval'92 | 4.94 | 3.60 | 5.03 |
| WSJ eval'93 | 6.94 | 4.98 | 8.08 |
| LibriSpeech test-clean | 7.89 | 5.33 | 5.83 |
| LibriSpeech test-other | 21.74 | 13.25 | 12.69 |

Figure: Comparison of WER for two speech systems and human level performance on **read** speech (from (Amodei et al., 2016))

| Accented Speech | | | |
|----------------------------|-------|-------|-------|
| Test set | DS1 | DS2 | Human |
| VoxForge American-Canadian | 15.01 | 7.55 | 4.85 |
| VoxForge Commonwealth | 28.46 | 13.56 | 8.15 |
| VoxForge European | 31.20 | 17.55 | 12.76 |
| VoxForge Indian | 45.35 | 22.44 | 22.15 |

Figure: Comparison of WER for two speech systems and human level performance on **accented** speech (from (Amodei et al., 2016))

On par with human transcription ?

| Noisy Speech | | | |
|------------------|-------|-------|-------|
| Test set | DS1 | DS2 | Human |
| CHiME eval clean | 6.30 | 3.34 | 3.46 |
| CHiME eval real | 67.94 | 21.79 | 11.84 |
| CHiME eval sim | 80.27 | 45.05 | 31.33 |

Figure: Comparison of WER for two speech systems and human level performance on **noisy** speech (from (Amodei et al., 2016))

Language coverage

- Google addresses (only) 100 languages (ASR)
- Language technology issues: 300 languages (95 % population)
- Language coverage / revitalisation / documentation issues: > 6000 languages !

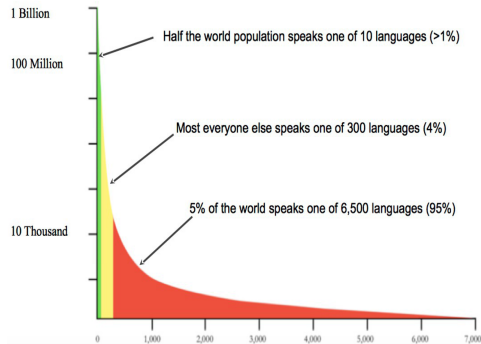


Figure: from Laura Welcher - Big Data for Small Languages The Rosetta Project

Low resource ASR

- Rapid development of ASR for new languages
- In low resource conditions
- For languages poorly described
- ex: DARPA Babel program

| Language | ML-24 | | CTC | | Attention | |
|-----------|-------|------|------|------|-----------|------|
| | XE | ST | XE | ST | XE | ST |
| Pashto | 54.5 | 51.5 | 51.5 | 49.5 | 50.6 | 50.1 |
| Guarani | 52.1 | 48.5 | -- | 47.8 | 47.3 | 46.3 |
| Igbo | 63.4 | 60.2 | 61.4 | 59.2 | 59.6 | 58.8 |
| Amharic | 49.3 | 44.5 | -- | 44.6 | 45.9 | 44.5 |
| Mongolian | 59.1 | 55.1 | 59.6 | 53.0 | 54.5 | 53.5 |
| Javanese | 60.1 | 55.3 | 57.5 | 54.1 | 55.4 | 54.2 |
| Dholuo | 43.7 | 40.5 | -- | 40.0 | 40.7 | 39.9 |
| Georgian | 45.0 | 40.8 | 44.3 | 40.6 | 45.6 | 43.9 |

Figure: Performance of end-2-end models on Babel languages⁵

⁵from Bhuvana Ramabhadran's presentation at Interspeech 2018

Zero resource ASR

In an unknown language, from unannotated raw speech, discover:⁶

- Invariant subword units (phone units ?)
- Words/terms (lexicon/semantic units ?)

Technological challenge

- Can we build useful speech technologies without any textual resources ?
- Unsupervised ASR / autonomous systems

Scientific challenge

- Can we build algorithms that learn languages like infants do ?
- Can we build algorithms that extract meaningful units from unknown languages ?

⁶The zero resource challenge: <http://zerospeech.com> (Dunbar et al., 2017)

Multilingual ASR

1 system - N languages

- In end-2-end ASR, acoustic, pronunciation and language model are integrated into a single neural network
- Makes them very suitable for truly multilingual ASR
- First attempts using hybrid CTC/attention ASR approaches (Watanabe et al., 2017)
- Similar in spirit to multilingual NMT (Johnson et al., 2016)
- Recent proposition using a transformer network (Zhou et al., 2018)

Other ASR challenges

- Can we leverage multiple sensors to design noise robust approaches ? (Barker et al., 2017)
- What do NNs learn ? (Belinkov and Glass, 2017)
- How can we can exploit adversarial examples to improve overall robustness ?
- Can we analyze (and deal with) biases between genders, dialects, regional accents ?
- How to deal with code-switching phenomena ?

Questions?

Thank you

References I

- Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Hannun, A. Y., Jun, B., Han, T., LeGresley, P., Li, X., Lin, L., Narang, S., Ng, A. Y., Ozair, S., Prenger, R., Qian, S., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Wang, C., Wang, Y., Wang, Z., Xiao, B., Xie, Y., Yogatama, D., Zhan, J., and Zhu, Z. (2016). Deep speech 2 : End-to-end speech recognition in english and mandarin. In Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, pages 173–182.
- Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (2017). Multi-microphone speech recognition in everyday environments. Computer Speech & Language, 46:386–387.
- Belinkov, Y. and Glass, J. R. (2017). Analyzing hidden representations in end-to-end automatic speech recognition systems. In NIPS, pages 2438–2448.
- Dunbar, E., Cao, X., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E. (2017). The zero resource speech challenge 2017. CoRR, abs/1712.04313.
- Graves, A., Fernández, S., Gomez, F. J., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In ICML, volume 148 of ACM International Conference Proceeding Series, pages 369–376. ACM.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2016). Google’s multilingual neural machine translation system: Enabling zero-shot translation. CoRR, abs/1611.04558.

References II

- Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In ICASSP, pages 5206–5210. IEEE.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.
- Watanabe, S., Hori, T., and Hershey, J. R. (2017). Language independent end-to-end architecture for joint language identification and speech recognition. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 265–271.
- Zhou, S., Xu, S., and Xu, B. (2018). Multilingual end-to-end speech recognition with A single transformer on low-resource languages. CoRR, abs/1806.05059.