Automatic Speech Recognition: Introduction, Current Trends and Open Problems

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)
- Can 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)

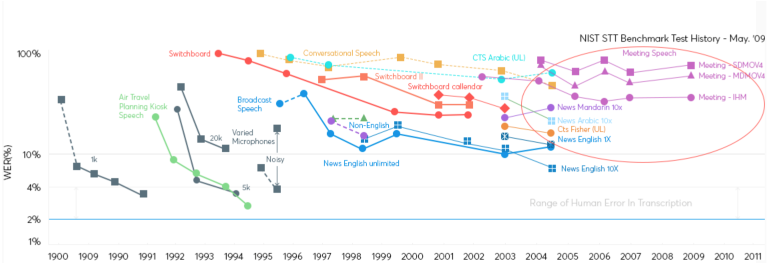


Figure: NIST ASR benchmark tests history (< 2015)

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

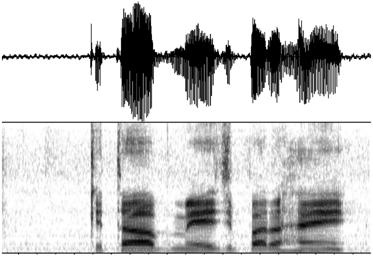


Figure: Speech signal (top) and spectrogram (bottom)

Progresses over the years

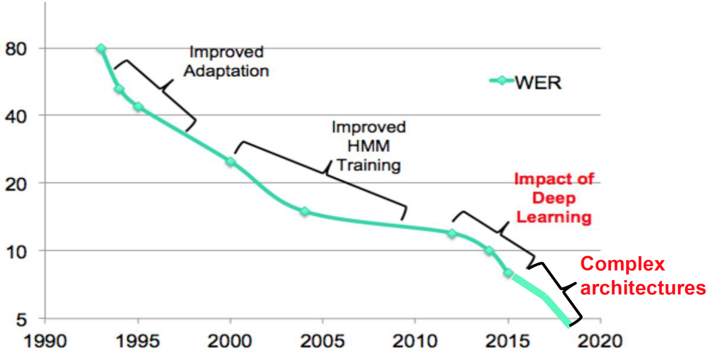


Figure: ASR Performance[1] on English Conversational Telephony (Switchboard)

[1] Image from Bhuvana Ramabhadran' s presentation at Interspeech 2018

Fundamental equation

x: observation (signal or features)
w: a word sequence

$$w^* = argmax_w p(w/x) = argmax_w p(x/w).p(w) \tag{1}$$

$p(x/w)$: 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

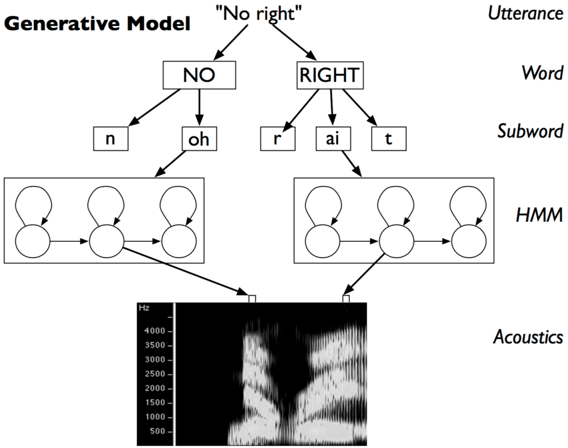


Figure: From speech to utterances[2]

[2] Image from Steve Renals's lecture on ASR

ASR overview

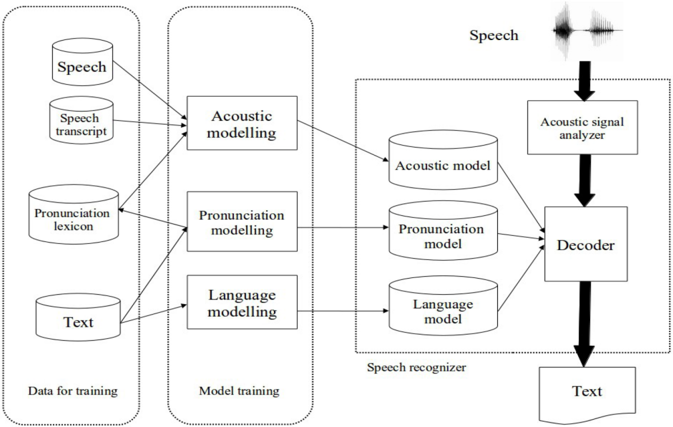


Figure: ASR Overview

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)

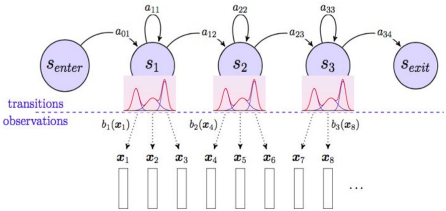


Figure: HMM/GMM approach

HMMs

- Well known algorithms for
  - *training* the model parameters (Baum-Welch algo.)
  - *decoding* the most probable hidden state sequence (Viterbi algo.)
  - *evaluate* the likelihood of an observation being generated by a HMM (Forward algo.)
- Phonemes are generally modeled in context (1 phoneme = N HMMs)
  - triphones or quintphones (model co-articulation)
  - state or parameter tying to reduce model complexity

GMMs

- Approximate a true distribution
- Easily trained with EM algorithm
- Well designed for speaker adaptation (shift the gaussians!)
  - almost 20 years of literature on speaker adaptation
- Tend to be replaced by DNNs since 2010
  - smaller footprint than GMMs
  - model several frames in a raw to increase context
  - speaker adaptation is an issue (less clear how to do it)

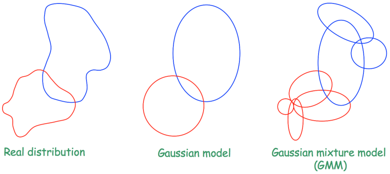


Figure: Gaussians in 2 dimensions

Language models: from N-grams to RNNs

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

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

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

n-gram LM: $h = w_{k-n+1}, w_{k-n+2}, ..., w_{k-1}$
recurrent neural network LM: $h = rnn\_state(E(w_1), E(w_2), ..., E(w_{k-1}))$

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

NNs for acoustic modeling (1990-2010)

- In most approaches, NNs model the posterior probability $p(s|x)$ of an HMM state s given an acoustic observation x
- Existing HMM speech recognizers can be used
- This model is known as hybrid NN-HMM and was introduced by Renals et al. (1994)

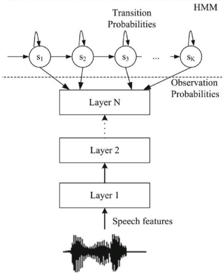


Figure: Hybrid NN-HMM

NNs for language modeling (1990-2010)

- Rescoring a lattice of output hypotheses using NN LM instead of N-gram
- Introduced by Bengio et al. (2003)
- Extended to large vocabulary speech recognition (Schwenk, 2007)
- Reducing computational complexity
  - using shortlist at output layer (Schwenk, 2007)
  - hierarchical decomposition of output probabilities (Morin and Bengio, 2005; Mnih and Hinton, 2008; Le et al., 2011)
- Recurrent neural networks were used in LM training (Mikolov et al., 2010)

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 !

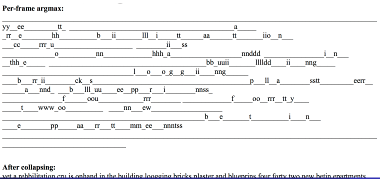
End-to-end ASR (get rid of HMMs)

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__ee_________tt_ ___________________________________a_____
_rr__e________hh__________b___ii_______lll__i_____tt______aa_______tt________iio__n___
___cc______rrr_u______________________ ________ii___ss
_______________o__________nn_____________hhh_a____________________nnddd _________________i__n___
__thh_e______ ____________________________________________bb_uuii_______llllldd____ii____nng_____
_____________________________________l___o____o_g__g____ii____nng_______
____b___rr_ii________ck__s____________________________________________p___ll__a________sstt_________eerr__
______a___nnd_ ___b___lll_uu_____ee__pp___r___i________nnss_
_________________f______oou_____________rrr_________ _______________f______oo__rrr__tt_y____
_____t____www_oo___________ ____nn___ew____________________
__________________________________________________________b___e_______t__________i____n___
____e_________pp______aa___rr___tt____mm_ee___nnntss
____________________________________________________________
___________________
```

After collapsing:

yet a rehbilitation cru is onhand in the building loogging bricks plaster and blueprins four forty two new betin epartments

Attention modeling

Initially proposed for (neural) machine translation (Bahdanau et al., 2014) and introduced for ASR by Chorowski et al. (2015)

- A context (attention) model is a function of the encoder codes and of the previous decoded tokens
- A speech encoder is defined (CNNs, pyramidal LSTMs)
- While CTC generates frame-level posteriors, attention models generate L predictions until the end-of-sequence symbol (no posterior for a given frame)
- Well-known issue with attention and CTC models is the thin lattices we end up with

Self attention for speech ?

- Recent paper at interspeech proposed self-attention for speech encoding (Sperber et al., 2018)
- 2 papers adapt transformer architecture to ASR (Zhou et al., 2018a,b)

Reading group ?

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 !

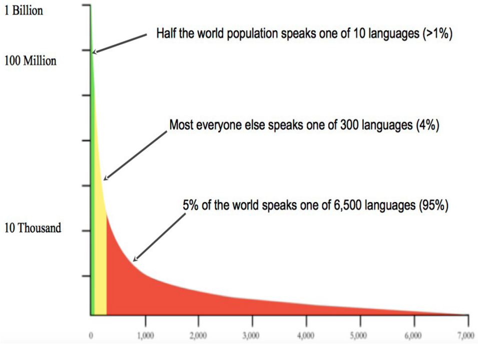


Figure: from Laura Welcher - Big Data for Small Languages - The Rosetta Project

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 languages[5]

[5] from 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 ?

Reading group ?

[6] The 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., 2018b)

Reading group ?

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 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, pages 173–182.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.

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.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. J. Mach. Learn. Res., 3:1137–1155.

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. CoRR, abs/1506.07503.

References II

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.

Le, H. S., Oparin, I., Allauzen, A., Gauvain, J., and Yvon, F. (2011). Structured output layer neural network language model. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic, pages 5524–5527.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In Interspeech.

Mnih, A. and Hinton, G. (2008). A scalable hierarchical distributed language model. In In NIPS.

Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In AISTATS05, pages 246–252.

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.

References III

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.

Renals, S., Morgan, N., Bourlard, H., Cohen, M., and Franco, H. (1994). Connectionist probability estimators in HMM speech recognition. IEEE Trans. Speech and Audio Processing, 2(1):161–174.

Schwenk, H. (2007). Continuous space language models. Computer Speech & Language, 21(3):492–518.

Sperber, M., Niehues, J., Neubig, G., Stüker, S., and Waibel, A. (2018). Self-attentional acoustic models. CoRR, abs/1803.09519.

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., Dong, L., Xu, S., and Xu, B. (2018a). Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese. In Interspeech, pages 791–795. ISCA.

Zhou, S., Xu, S., and Xu, B. (2018b). Multilingual end-to-end speech recognition with A single transformer on low-resource languages. CoRR, abs/1806.05059.