Prediction of non-linguistic speaker attributes from voice

Dec 16, 2021

The MAMA AI

Outline

- Introduction/Motivation
- EXTRA TOPIC: Beyond “traditional” HMM ASR - RNN-T ASR approach
- Acoustic features of human voice (connection to previous talks on Biometry)
- Non-linguistic features
- ML pipeline
  - Data collection (incl. Cleaning, labeling)
  - Feature extraction
  - Model training and evaluation
  - API/Deployment
    - Kubernetes stack, websocket API
  - GUI for demo purposes
- EXTRA TOPIC: Text-to-Speech Synthesis - state of the art system from our company
  - Architecture
  - Model training
  - Demo voices

About us

World class AI and Cloud experts with 25+ years of experience

Trustworthy relationships with customers and industry partners

IBM Watson and IBM Research alumni

Multigenerational team, balanced and highly-performing

Mama AI team "stats"

The MAMA AI

270	Years in IBM
8	Years in Startups
51	Years in Academia
165	Years in Speech
29	Years in Machine Translation
125	Years in Dialog
80	Years in Neural Networks, Deep Learning
125	Years in Statistical NLP
118	Years in NLP
11.5	Years in SRE
42	Years in k8s
33	Years in IoT

302	Publications
6036	Citations
95	Patent applications filed

15521	Miles run to date
1	Ironmans completed

1	accoredonist
1.6	ukulelist
4.8	pianist
4.5	guitar
1	bass
1	violin
0.7	drums
27425	bullet chess games
250	floorball matches
53	Sněžka summit
33	Říp summit
230	countries visited

End-to-end ASR - Conventional ASR

-

$$P(T|A)$$

T ... text, A ... acoustics

- Unable to model P(T|A) directly, so using Bayes:

$$P(T|A) = \frac{P(A|T)P(T)}{P(A)}$$

  - P(A) constant, ignoring
  - P(A|T) ... acoustic model P_A
  - P(T) ... language model P_T

Acoustic model $P_A(A|T)$

-
- T... hypothesized sequence of acoustic units
- we assume independence between frames so that we can write

$$P(A|T) = \prod_i P_A(a_i\,|t_i) \qquad \text{i ... time}$$

- i ... time
- a_i ... feature vector
- t_i ... acoustic class (a phone or context-dependent phone)

Language model $P_{LM}(T)$

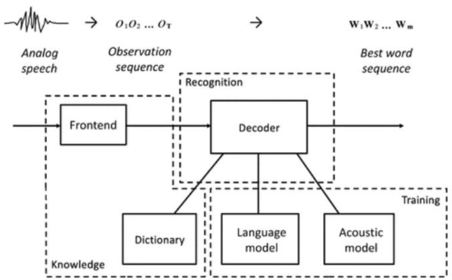
-
- T... sentence, consists of word sequence w_1 ... w_N
- sequence probability modelled with n-gram LM (or neural LMs)

$$P_{LM}(T) = \prod_i P(w_i\,|w_{i-1}, w_{i-2}, \dots)$$

How to map words w_i to acoustic classes t_i ? → need Dictionary $P_{pron}(pron|w)$

- "several" S EH V AX R AX L
- "several" S EH V R AX L

Decoder



End-to-end ASR - Conventional ASR

- $$P(T|A)$$

  T ... text, A ... acoustics

- Unable to model P(T|A) directly, so using Bayes:

$$P(T|A) = \frac{P(A|T)P(T)}{P(A)}$$

  - P(A) constant, ignoring
  - P(A|T) ... acoustic model P_A
  - P(T) ... language model P_T

End-to-end ASR - Conventional ASR

Acoustic model $P_A(A|T)$

- T... hypothesized sequence of acoustic units
- we assume independence between frames so that we can write

$$P(A|T) = \prod_i P_A(a_i \,|t_i) \qquad \text{i ... time}$$

- i ... time
- a_i ... feature vector
- t_i ... acoustic class (a phone or context-dependent phone)

End-to-end ASR - Conventional ASR

Language model $P_{LM}(T)$

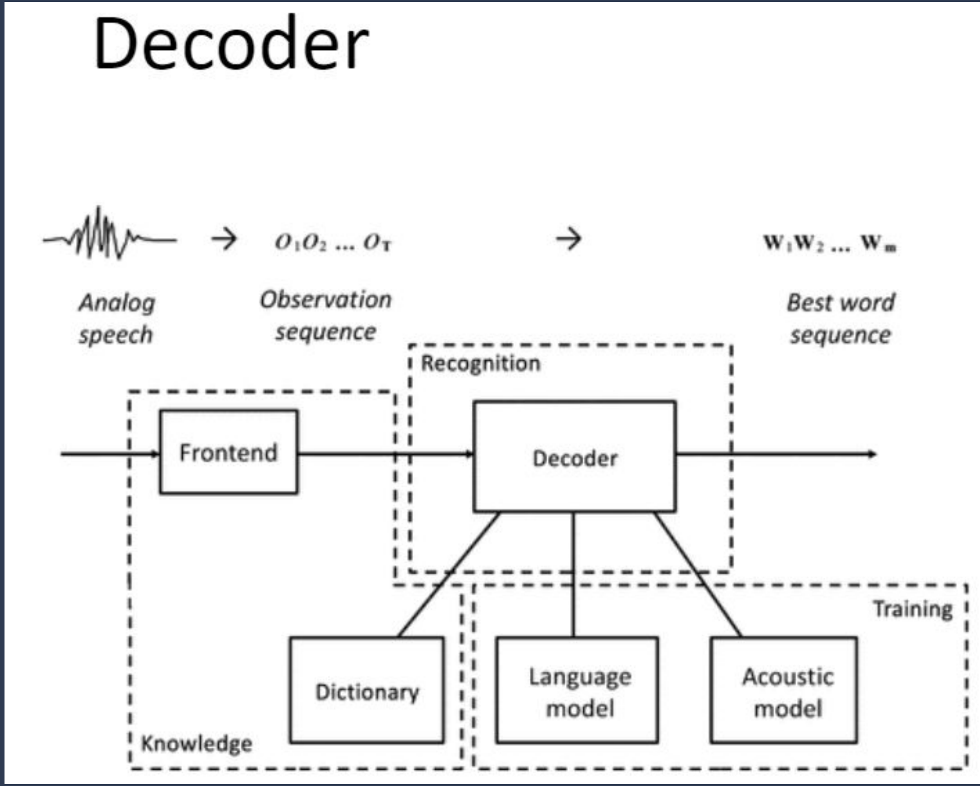
-
- T… sentence, consists of word sequence w_1 … w_N
- sequence probability modelled with n-gram LM (or neural LMs)

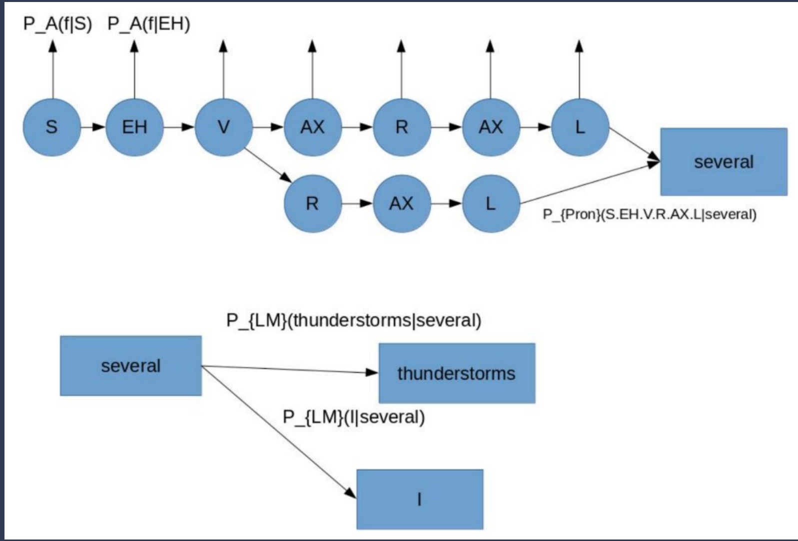
$$P_{LM}(T) = \prod_i P(w_i \,|w_{i-1}, w_{i-2}, \dots)$$

How to map words w_i to acoustic classes t_i ? → need Dictionary $P_{pron}(pron|w)$

- "several" S EH V AX R AX L
- "several" S EH V R AX L

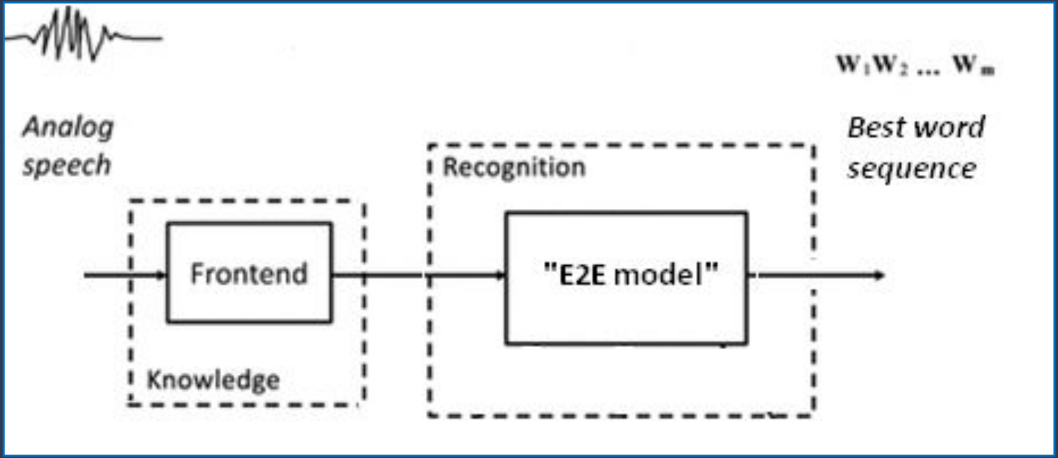
End-to-end ASR - Conventional ASR



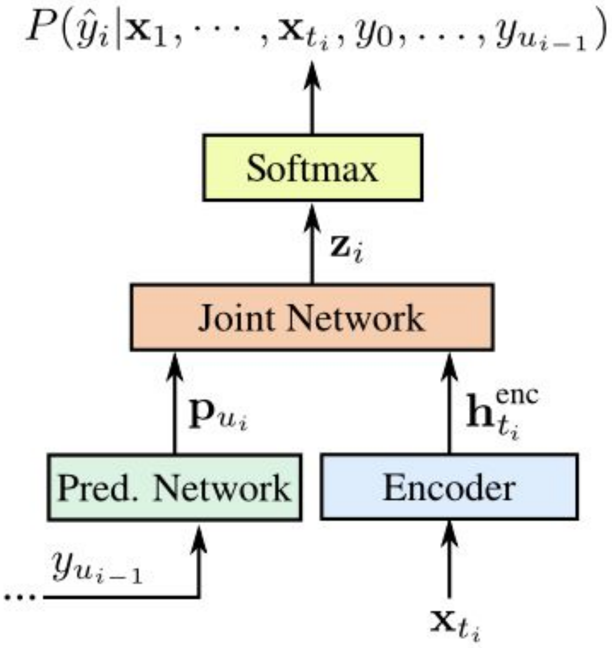


End-to-end ASR (E2E)

- Models $P(T|A)$ directly
- Able to predict per-frame probability of characters, sub-words or even words
- No need for Dictionary (pronunciation is not modelled explicitly)
- No separate acoustic model → no need for alignment between symbols and audio
- Decoder can be much simpler
- All you need to build the model is audio and its transcript



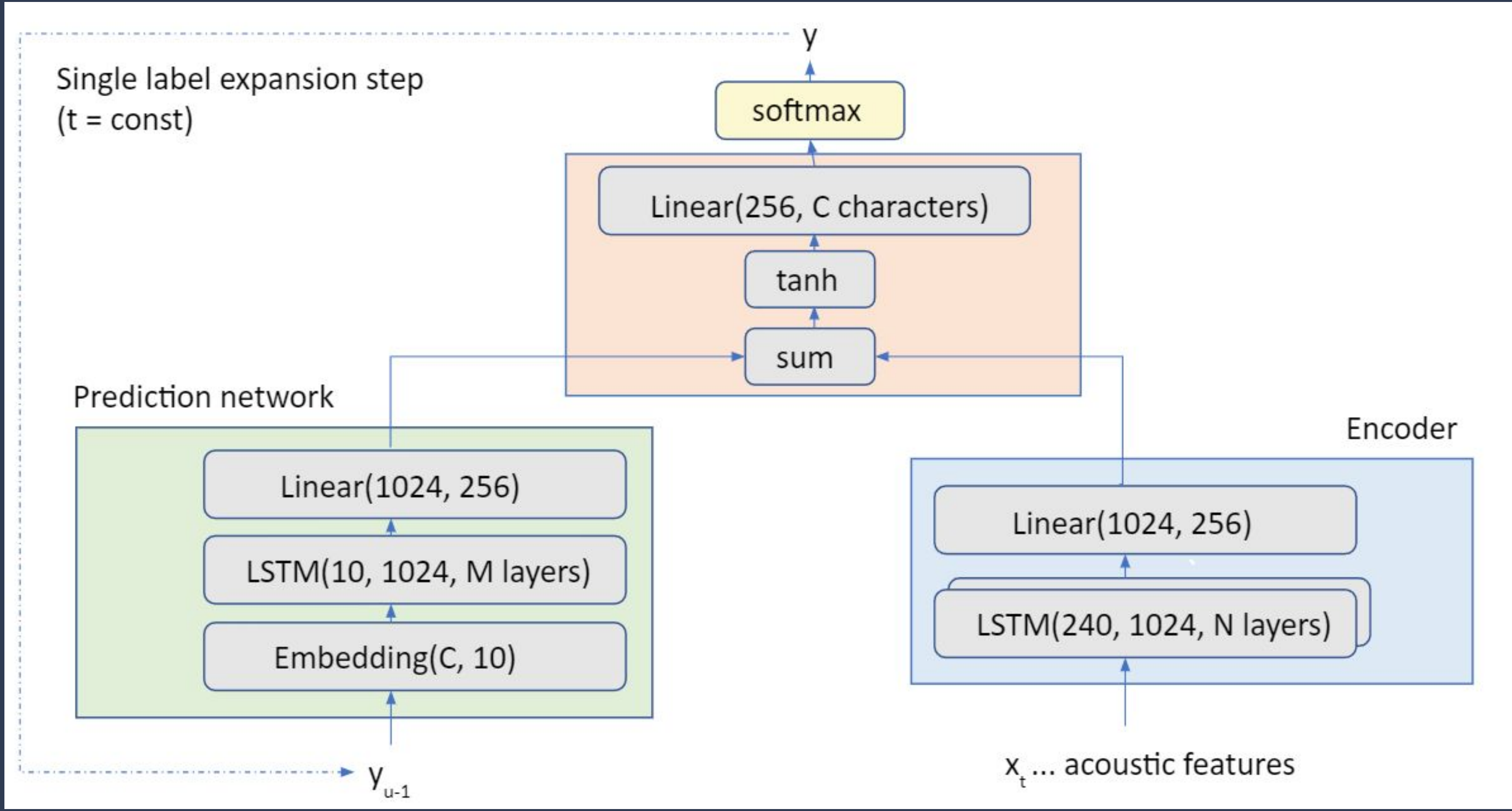
End-to-end ASR - RNN-Transducer



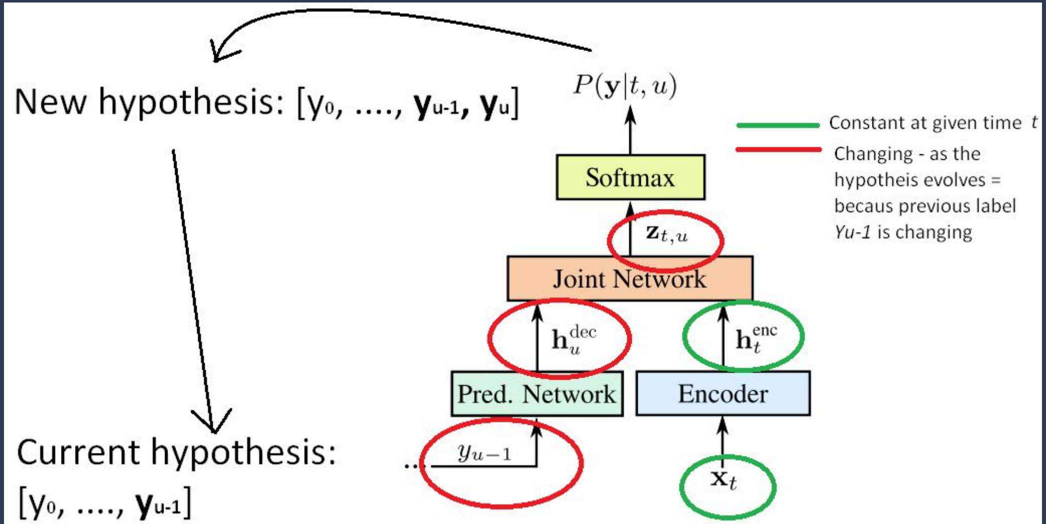
(b.) RNN-Transducer

* "Streaming End-to-end Speech Recognition for Mobile Devices", ICASSSP'19, Google

End-to-end ASR - RNN-T model architecture – an example



End-to-end ASR - RNN-T Decoder



- Beam search
- RNN-T gives P(y|x,y_), where y is the next character, x are the audio frames so far, y_ is the current hypothesis
- RNN-T does not always consume input (allows to decode multiple characters in a single frame)

End-to-end ASR - RNN-T loss

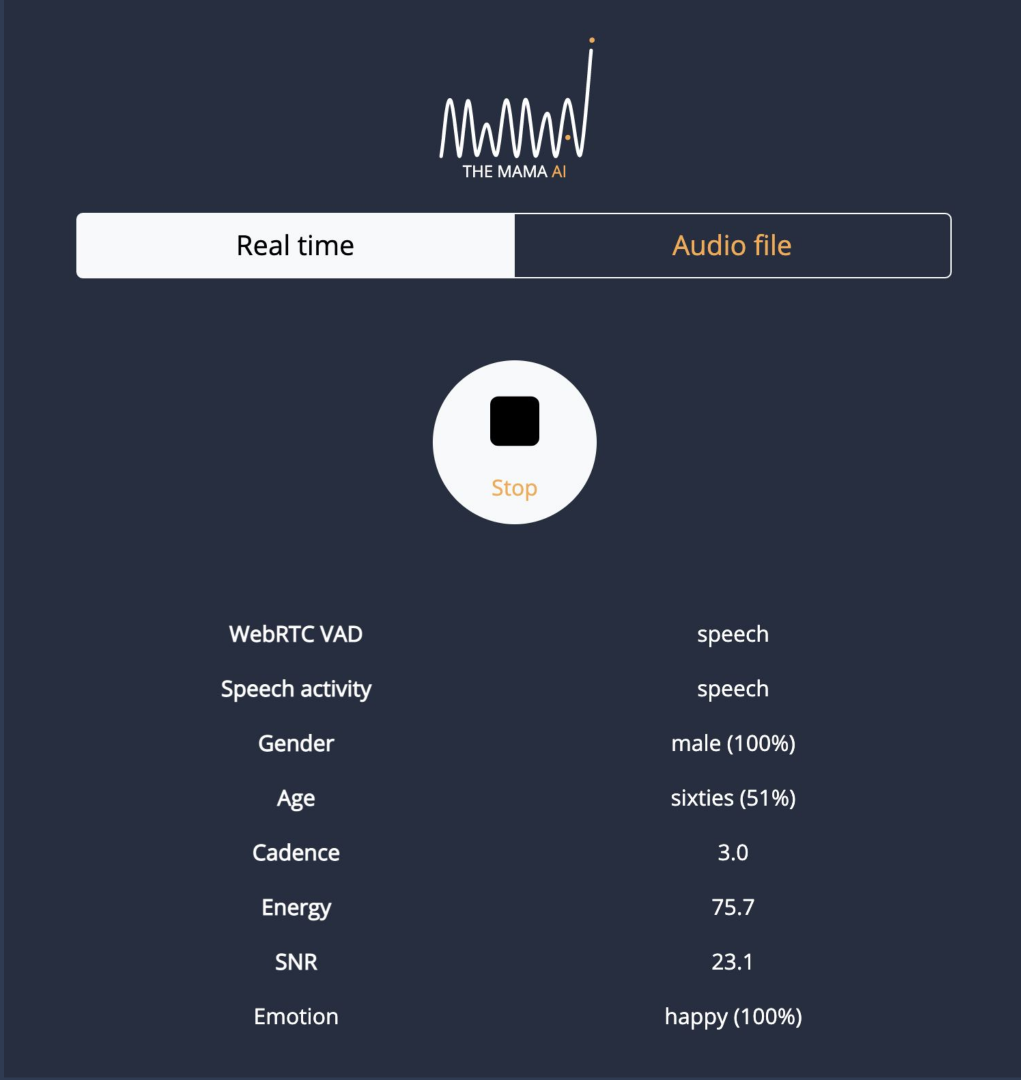
- For each frame, NN outputs probabilities of characters + blank symbol
- We have the correct transcript in train time
- An alignment is a sequence of characters + blank symbols
- A consistent alignment is one consistent with the correct transcript, e.g.
- Let's say we have 8 acoustic frames and a transcript "hello”
- an alignment "_ _ h e _ l _ l _ o _ _ _" is consistent
- so is " _ h _ e _ _ l l o _ _ _ _ _”
- by definition, blank symbol means go to next frame
- RNN-T optimizes the sum of probabilities across all consistent alignments

Introduction/Motivation
Biospeech engines

- Dialog systems
  - Speech Activity - spare speech recognition cycles when no one speaks
  - Gender - verb forms “Co byste nám chtěl**(a)** sdělit”
  - Speaking rate - matching speaking rate improver user experience
  - Age - adapt speaking style/tempo for different age cohorts
  - SNR - adapt to level of noise
    - use confirmations more often in noisy condition
    - Be more “greedy” when noise level is low
    - ask the user to speak louder or call from a quieter place
  - Emotion
    - Hand-over to human agent if customer gets too angry
    - Analyze interaction patterns with respect to detected emotions

Biospeech

- **Leveraging signals in spoken word to identify number of biomarkers**
    - **understand mood** - to be able to approach clients differently
    - **measure stress** during tasks and interactions
    - measure and help control rate of person's speech etc.
    - recognize **gender and age** group based on voice



The non-linguistic speaker attributes from voice - Demo Page

Current version of web demonstrates detection of:

- Speech activity (WebRCT and Neural)
- Gender
- Age in decades
- Cadence Praat based and Neural (syllables/second)
- Energy of signal
- Signal to noise ratio SNR in dB
- One of four emotions (angry, happy, sad, neutral)

RECURRENT NEURAL NETWORKS WITH LOCAL ATTENTION for Emotion, Gender, Age:

Original paper uses Low level descriptors based on the Praat framework:

- pitch (F0),
- voicing probability,
- energy,
- zero-crossing rate,
- Mel-filterbank features,
- MFCCs,
- formant locations/bandwidths, harmonics-to-noise ratio, jitter, etc.).

Our version uses log-mel features

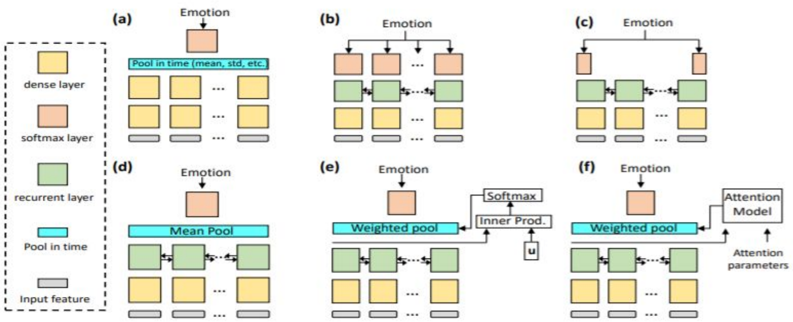
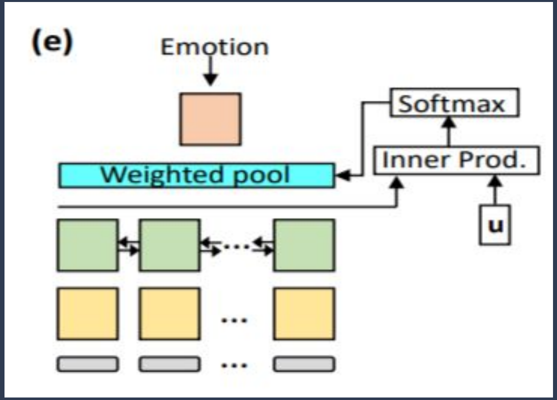
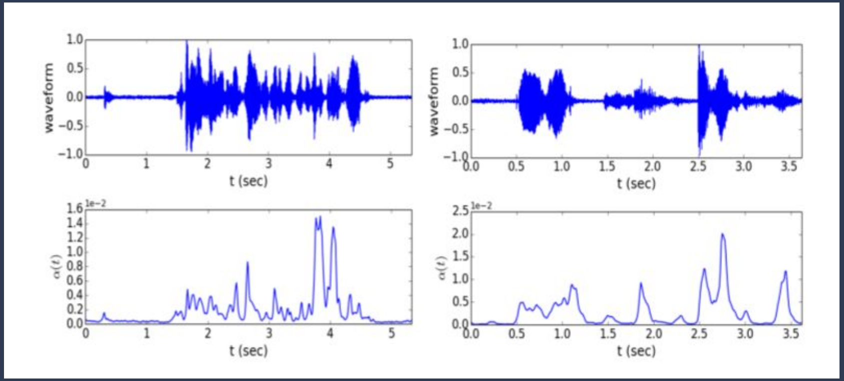


Fig. 1. Architectures for applying DNN/RNN for SER. (a) Learning LLDs using fixed temporal aggregation. (b) frame-wise training. (c) final-frame (many-to-one) training. (d) Mean-pooling in time. (e) Weighted pooling with logistic regression attention model. (f) general attention model.

AUTOMATIC SPEECH EMOTION RECOGNITION USING RECURRENT NEURAL NETWORKS WITH LOCAL ATTENTION Seyedmahdad Mirsamadi1, Emad Barsoum2,

The non-linguistic speaker attributes from voice - Weighted pooling with logistic regression





$$\alpha_t = \frac{\exp(\mathbf{u}^T \mathbf{y}_t)}{\sum_{\tau=1}^{T} \exp(\mathbf{u}^T \mathbf{y}_\tau)}.$$

AUTOMATIC SPEECH EMOTION RECOGNITION USING RECURRENT NEURAL NETWORKS WITH LOCAL ATTENTION Seyedmahdad Mirsamadi1, Emad Barsoum2,

The non-linguistic speaker attributes from voice - Log-Mel features

Signal -> [Hamming] -> [FFT] -> [abs()^2] -> [x*Mel] -> [log()] ->log-mel

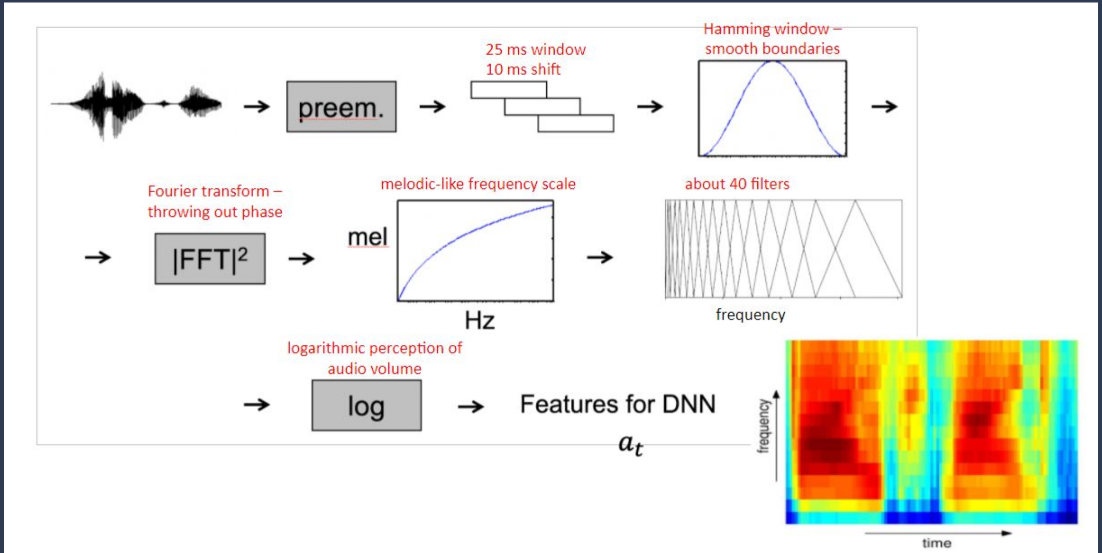
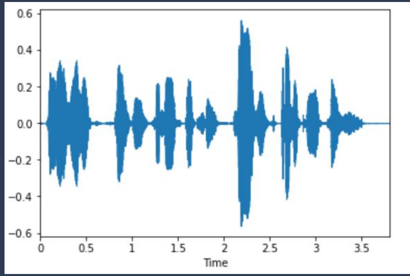
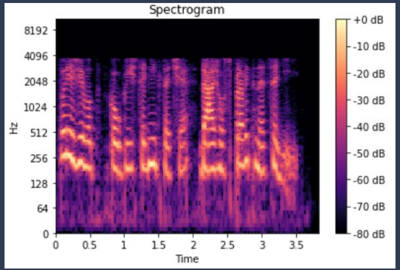
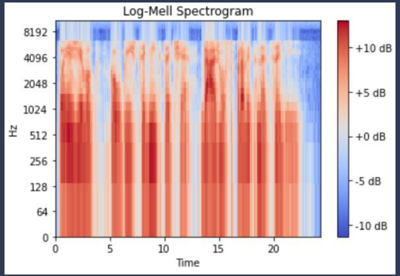


Image from MathWorks







The non-linguistic speaker attributes from voice - Articulation rate / Speech cadence

- Bidirectional LSTM
  - Input dimension 40
  - hidden dimensions 300
  - 3 layers
- Fully connected linear layer → dimension 1
- Mean over all LSTM steps
- MAE: 0.328 syll/sec
  - More than double the performance of Praat (0.767 syll/sec)
- Speed: ~100x faster than real time

The non-linguistic speaker attributes from voice - Energy, SNR

- Energy P_{signal} of speech is measured in the speech segments of the audio
- SNR is computed from energy of speech P_{signal} and background noise measured in non-speech segments P_{noise}
- For identification of speech/non-speech segments the speech detectors or the Praat framework can be used.

$$\mathrm{SNR_{dB}} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right).$$

MAMA AI Text-to-Speech

Motivation (Why do it on our own?)

- It is not only the latest tech
- Training pipeline
  - Data preprocessing and cleaning
  - Data validation
  - Custom voices (incl. recording)
- Model management and provisioning
- APIs for serving the model
- Deployments beyond Cloud
  - Variability in access and pricing per customer needs

MAMA AI Text-to-Speech

Architecture

- Data segmentation
- Text normalization
- Phonetization
- Forced alignment
- mel-spectrogram generator FastPitch
- Vocoder HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis

Samples:

We *are* passionate about AI

The MAMA AI

Conversational AI

Assistants and chatbots,
design for voice and text,
deflection of common tasks,
customer sentiment

Speech

Call center transcriptions,
call logs analytics,
speaker id and verification,
mood detection, agent guidance

Edge/IoT/Hybrid

NLP and Voice on
embedded platforms
(gaming, automotive,
remote/offline use)

Omni-Channel Interaction

Interactive customer
notifications,
upsell/cross-sell,
user profiles

Natural Language Processing

Natural Language
Understanding, written
or spoken reports from
structured data

Applied AI

Acoustic monitoring
and prediction,
predictive maintenance,
AIOps

The MAMA AI

https://themama.ai
mama@themama.ai

The non-linguistic speaker attributes from voice - BACKUP

- Apha
- Beta

The non-linguistic speaker attributes from voice - Log-Mel features

Signal -> [Hamming] -> [FFT] -> [abs()^2] -> [x*Mel] -> [log()] ->log-mel