Deep Learning in Speech Recognition

Alex Acero
Siri

A brief history of Artificial Intelligence and Machine Learning

Birth of Artificial Intelligence

1950: Alan Turing creates the “Turing Test” to determine if a computer has real intelligence

1955: John McCarthy, Marvin Minsky, Nathaniel Rochester, Claude Shannon, Ray Solomonoff, Oliver Selfridge, Trenchard More, Arthur Samuel, Herbert Simon, and Allen Newell

> We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

Checkers

(Arthur Samuel, 1956)



ELIZA

(Weizenbaum,1966)

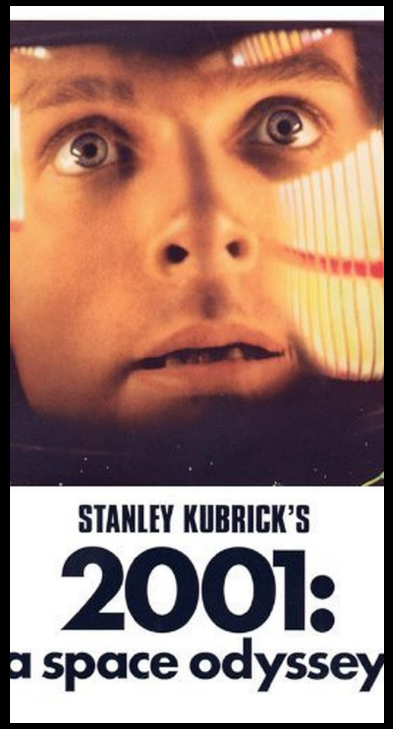
```
================================================================================
          EEEEEEEE  L          IIIIIII  ZZZZZZZ      AAA
          E         L             I           Z      A   A
          E         L             I          Z      A     A
          EEEEE     L             I         Z       A     A
          E         L             I        Z        AAAAAAA
          E         L             I       Z         A     A
          EEEEEEEE  LLLLLLLL  IIIIIII  ZZZZZZ      A     A

================================================================================
ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA. WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE... !
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...

================================================================================
```

2001 Space Odyssey

(Stanley Kubrick, 1968)



AI Winter



Deep Blue

(IBM, 1997)

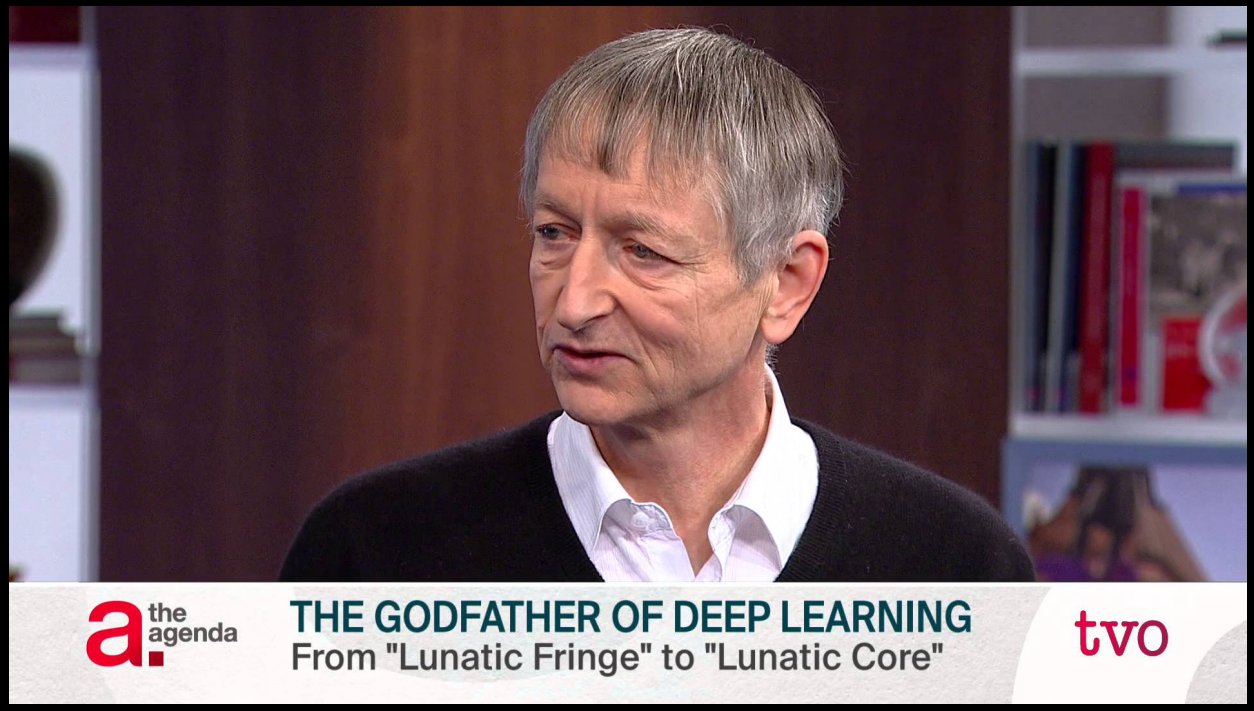


Autonomous driving

(Stanford, 2005)



Deep Learning
(Hinton, 2006)

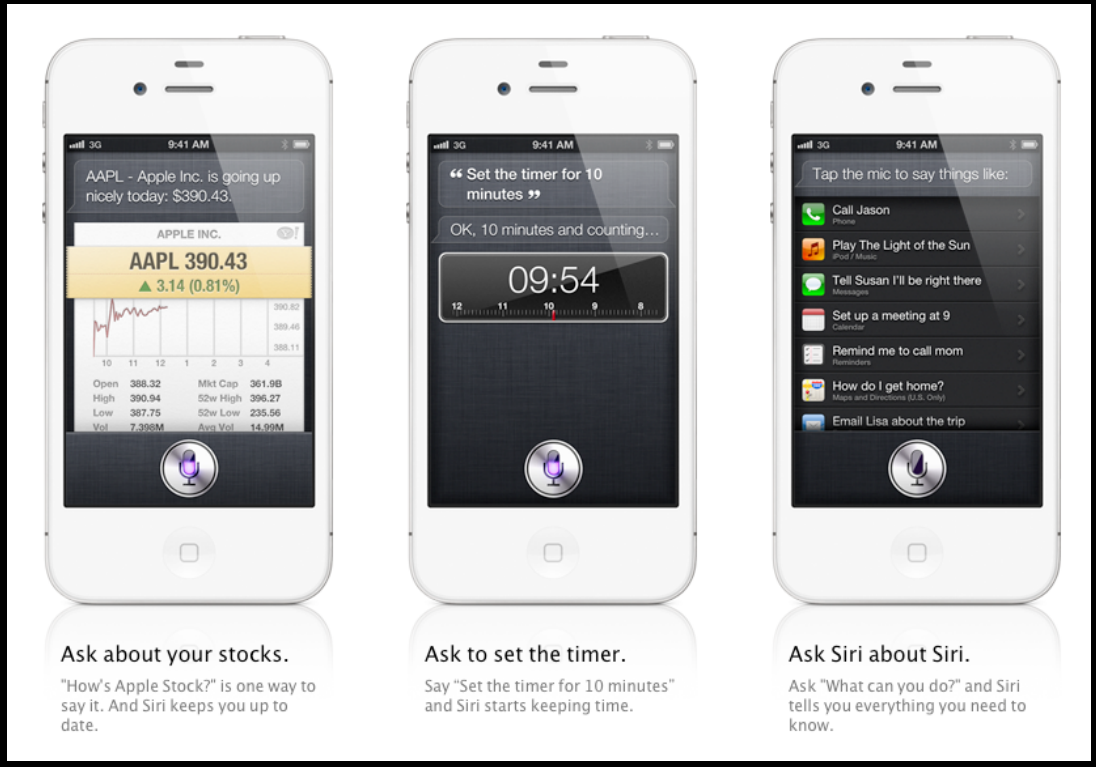


Jeopardy
(IBM, 2011)



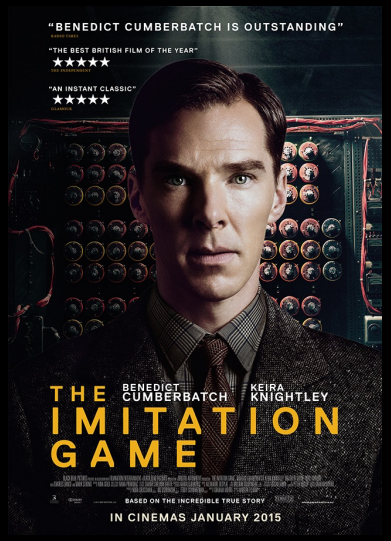
Siri

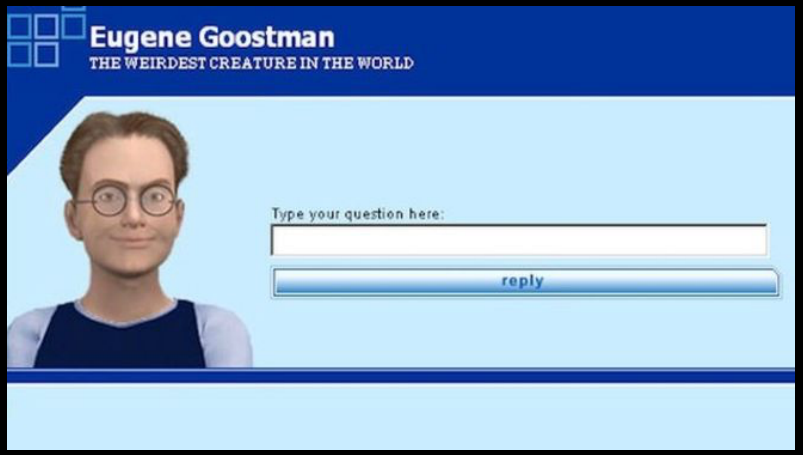
(Apple, 2011)



The imitation game

(2014)





Vladimir Veselov and Eugene Demchenko

The 65-year-old Turing Test is successfully passed if a computer is mistaken for a human more than 30% of the time during a series of five-minute keyboard conversations.

On 7 June, 2014 a computer program called Eugene Goostman (simulating a 13-year-old Ukrainian boy) convinced 33% of the judges at the Royal Society in London that it was human

AlphaGo

(Deepmind, 2016)



Introduction to Deep Learning

Machine Learning

Arthur Samuel (1959): Field of Study that gives Computers the ability to learn without being explicitly programmed

Tom Mitchell (1998): Improve on Task T, with respect to performance metric P, based on experience E



Improve on Task T with respect to performance metric P based on experience E

T: Categorize email messages as spam or legitimate
P: Percentage of email messages correctly classified
E: Database of emails, some with human-given labels



T: Recognizing hand-written words
P: Percentage of words correctly classified
E: Database of human-labeled images of handwritten words

T: playing checkers
P: percentage of games won against an arbitrary opponent
E: Playing practice games against itself

T: Driving on four-lane highways using vision sensors
P: Average distance traveled before a human-judged error
E: A sequence of images and steering commands recorded while observing a human driver

Binary Classification

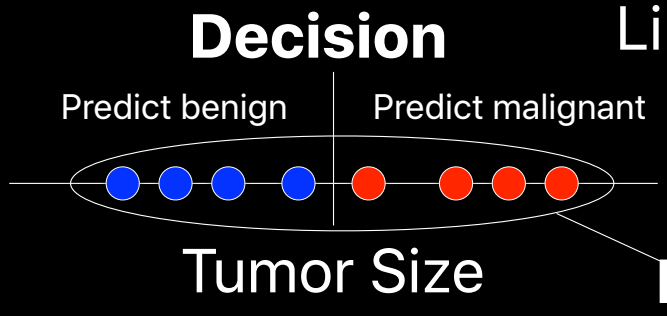
Output Labels

Breast Cancer

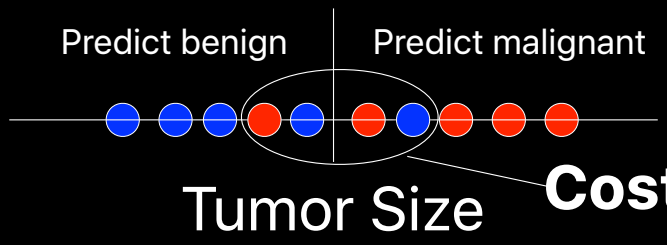
- Benign
- Malignant

Input Features

Tumor Size



Linearly Separable



Not Linearly Separable

Binary Classification

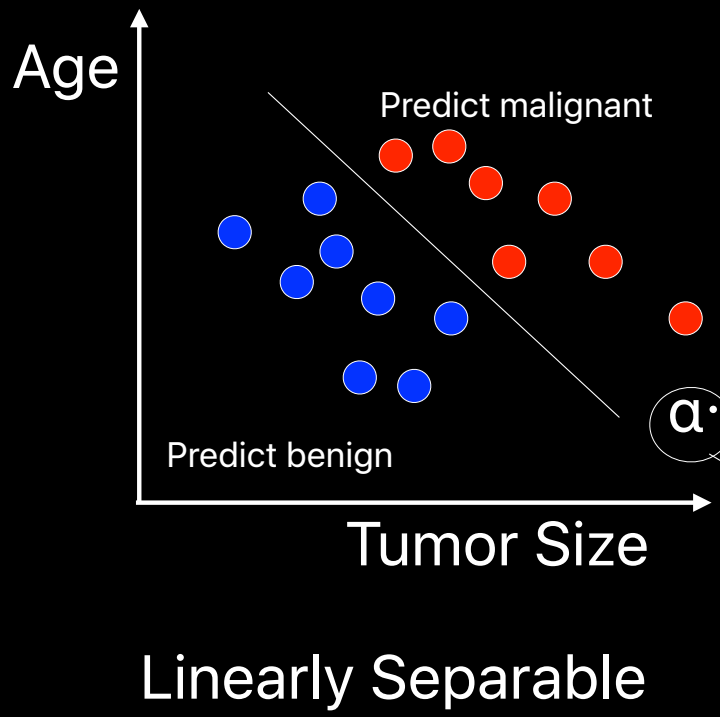
Output Labels

Breast Cancer

- Benign
- Malignant

Input Features

Tumor Size
Age



$\alpha \cdot \text{Age} + \beta \cdot \text{Size} + \lambda > 0$

Learning

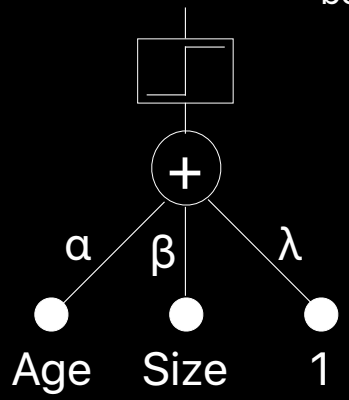
Linearly Separable

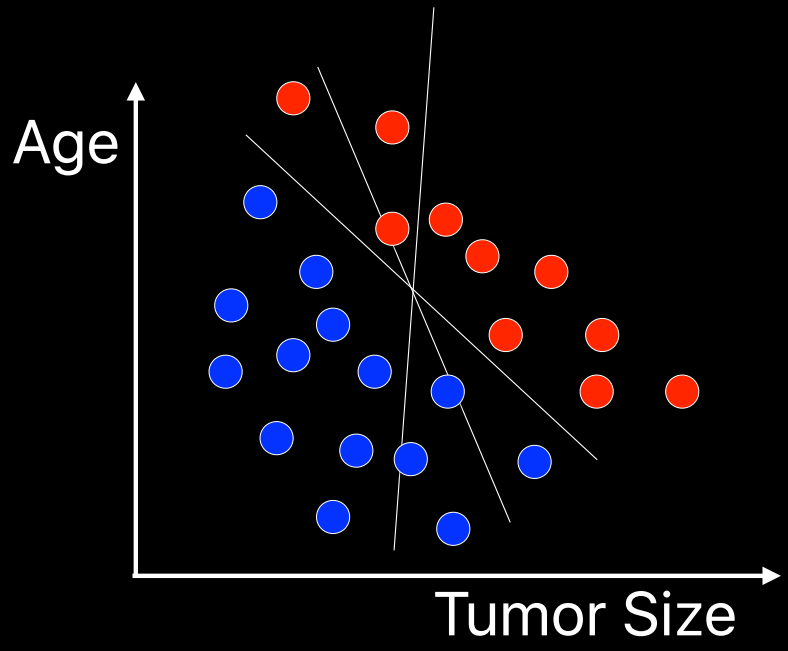
More features:
Clump thickness
Uniformity of cell size
...

Perceptron Learning

(Rosenblatt, 1957)

$$\alpha \cdot \text{Age} + \beta \cdot \text{Size} + \lambda \underset{\text{benign}}{\overset{\text{malignant}}{\gtrless}} 0$$

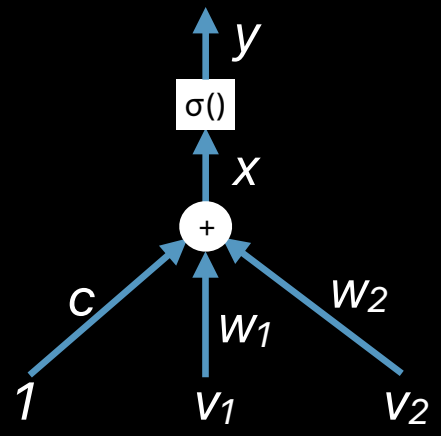




$$\alpha(i) = \alpha(i-1) + \eta \cdot \{\text{Target}(i) - \text{Output}(i)\} \cdot \text{Age}(i)$$
$$\beta(i) = \beta(i-1) + \eta \cdot \{\text{Target}(i) - \text{Output}(i)\} \cdot \text{Size}(i)$$
$$\lambda(i) = \lambda(i-1) + \eta \cdot \{\text{Target}(i) - \text{Output}(i)\}$$

A probabilistic framework

$$P(h, \mathbf{v}) = \frac{e^{\mathbf{b}^T\mathbf{v} + ch + \mathbf{v}^T\mathbf{w}h}}{Z}$$

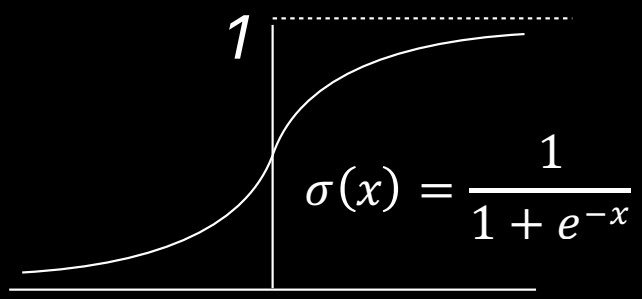


h is a binary variable

$$P(h|\mathbf{v}) = \frac{P(h, \mathbf{v})}{P(\mathbf{v})} = \frac{P(h, \mathbf{v})}{P(0, \mathbf{v}) + P(1, \mathbf{v})} = \frac{e^{\mathbf{b}^T\mathbf{v} + ch + \mathbf{v}^T\mathbf{w}h}}{e^{\mathbf{b}^T\mathbf{v}} + e^{\mathbf{b}^T\mathbf{v} + c + \mathbf{v}^T\mathbf{w}}} = \frac{e^{(c + \mathbf{v}^T\mathbf{w})h}}{1 + e^{(c + \mathbf{v}^T\mathbf{w})}}$$

$$y = P(h = 1|\mathbf{v}) = \frac{e^{c + \mathbf{v}^T\mathbf{w}}}{1 + e^{c + \mathbf{v}^T\mathbf{w}}} = \frac{1}{1 + e^{-(c + \mathbf{v}^T\mathbf{w})}} = \sigma(x)$$

$$x = c + \mathbf{v}^T\mathbf{w}$$



Loss function

Loss function between two probability distributions

$$D(p,q) = -\sum_{k=1}^{K} q(k) \ln p(k)$$

KL divergence or Cross-Entropy

For binary variables:

$$D = -t \ln y - (1-t) \ln(1-y)$$

$$\frac{\partial D}{\partial y} = \frac{-t}{y} + \frac{(1-t)}{1-y} = \frac{(y-t)}{y(1-y)}$$

$$y = \sigma(x) = \frac{1}{1+e^{-x}}$$

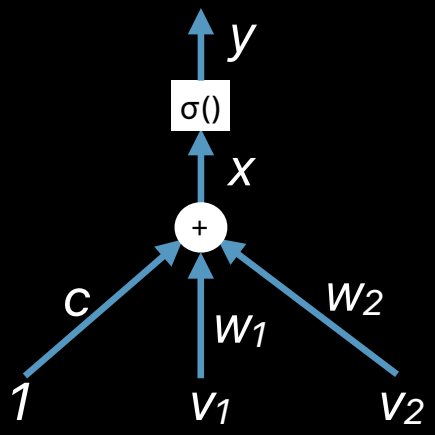
$$\frac{\partial y}{\partial x} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{(1+e^{-x})} \frac{e^{-x}}{(1+e^{-x})} = y(1-y)$$

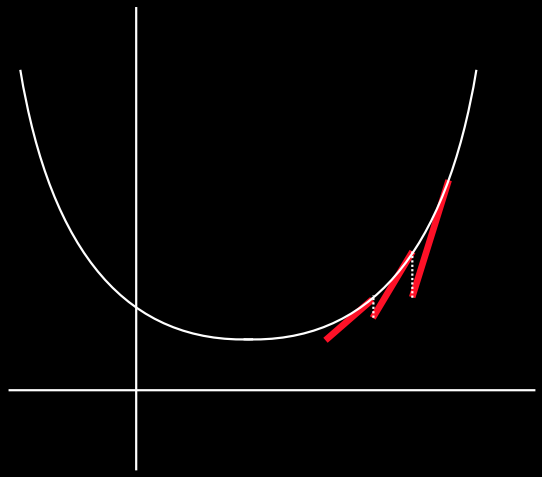
$$\frac{\partial D}{\partial x} = \frac{\partial D}{\partial y} \frac{\partial y}{\partial x} = y - t$$

Stochastic gradient descent

$$y = P(h = 1|\mathbf{v}) = \sigma(c + \mathbf{v}^T\mathbf{w})$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$





$$w_j^{(t)} = w_j^{(t-1)} - \eta\frac{\partial D}{\partial w_j} = w_j^{(t-1)} + \eta v_j\left(t - y^{(t-1)}\right)$$

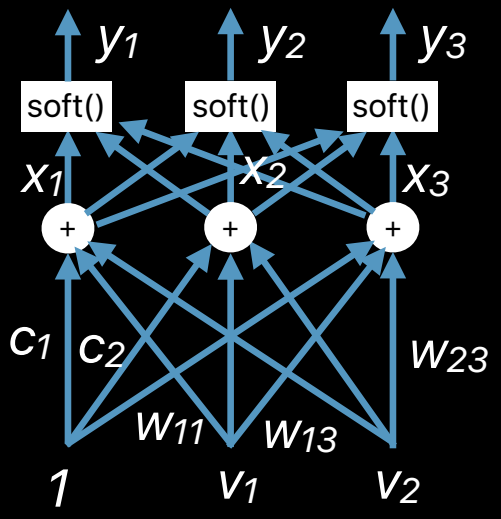
N-ary classification

$$D = -\sum_{i=1}^{N} t_i \ln y_i \qquad \frac{\partial D}{\partial y_i} = -\frac{t_i}{y_i}$$

$$y_i = p(i|\mathbf{v}) = \frac{e^{x_i}}{\sum_{l=1}^{N} e^{x_l}} \qquad \frac{\partial y_i}{\partial x_j} = -\frac{e^{x_i} e^{x_j}}{(\sum_{l=1}^{N} e^{x_l})^2} = -y_i y_j$$

Softmax

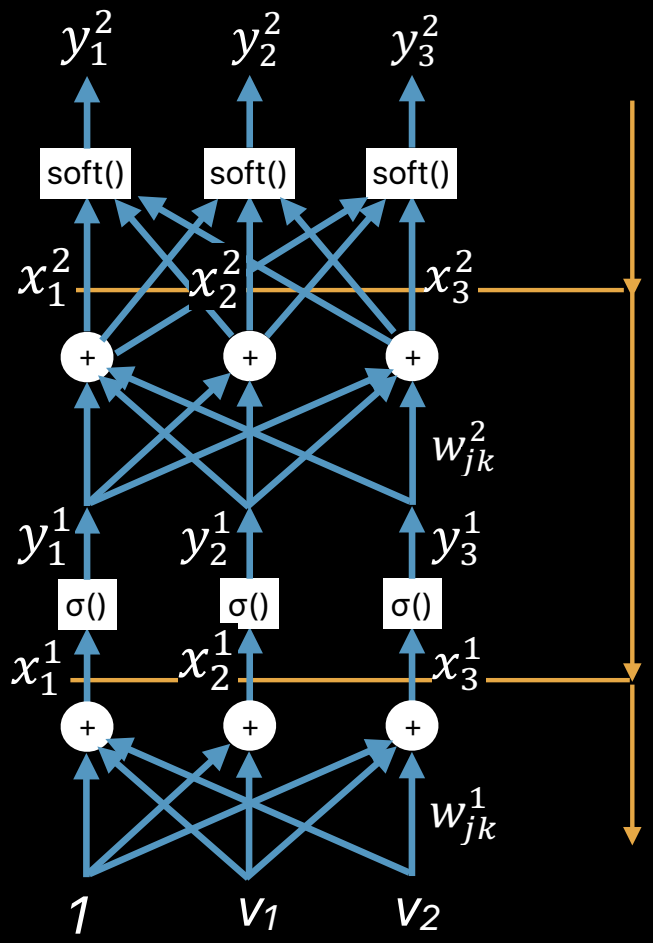
$$\frac{\partial y_i}{\partial x_i} = \frac{e^{x_i}}{(\sum_{l=1}^{N} e^{x_l})} - \frac{(e^{x_i})^2}{(\sum_{l=1}^{N} e^{x_l})^2} = y_i(1 - y_i)$$



$$\frac{\partial D}{\partial x_j} = \sum_{i=1}^{N} \frac{\partial D}{\partial y_i} \frac{\partial y_i}{\partial x_j} = -t_j + \sum_{i=1}^{N} \frac{t_i}{y_i} y_i y_j = -t_j + y_j \sum_{i=1}^{N} t_i = y_j - t_j$$

$$w_{ij}^{(t)} = w_{ij}^{(t-1)} + \eta v_i^{(t-1)} \left(t_j - \mathrm{y}_j^{(t-1)} \right)$$

Multi-layer Perceptron (Werbos, 1974)



Two-layers
2 input features
3 output labels

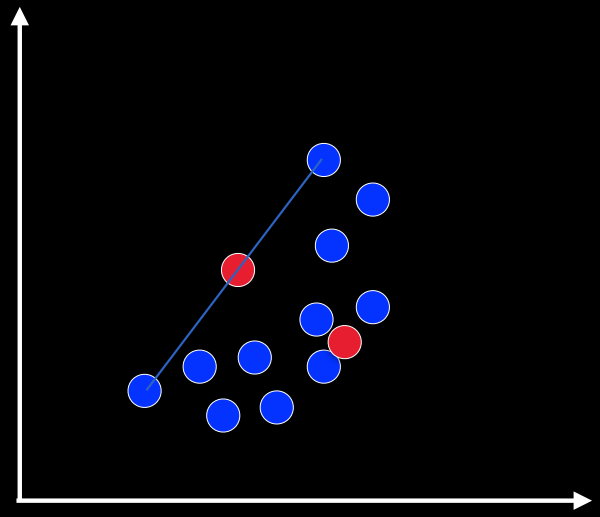
$$\nabla_n^2(m) = y_n^2(m) - t_n(m)$$

$$\nabla_n^1(m) = y_n^1(m)\left(1 - y_n^1(m)\right) \sum_{k=1}^{N} w_{nk}^2 \nabla_k^2(m)$$

$$\left[w_{jn}^1\right]^{(t)} = \left[w_{jn}^1\right]^{(t-1)} - \eta \sum_{m=1}^{M} v_j(m) \nabla_n^1(m)$$

backpropagation
Mini-batch

Non-linear manifolds



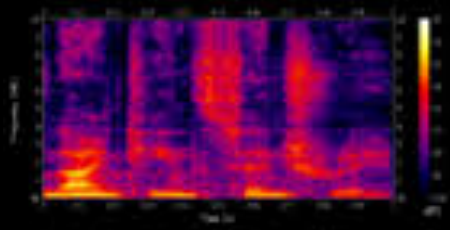
such nonlinearity
requires multiple layers

Binary Classification Tasks

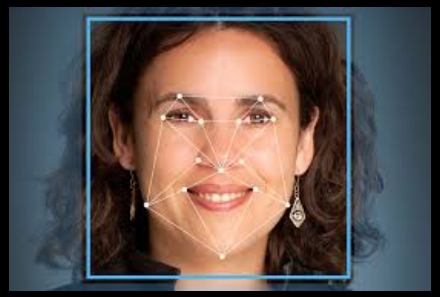
TouchID



Speaker Verification



Face ID



Email spam



Motion Detection



Credit Card fraud

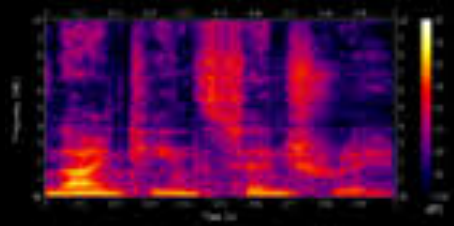


N-ary Classification Tasks

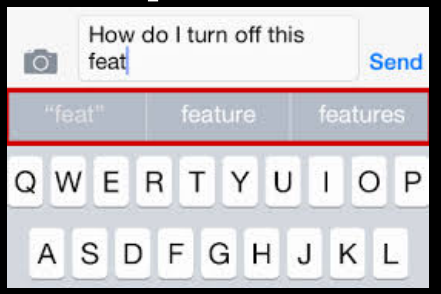
Handwriting Recognition



Speaker Identification

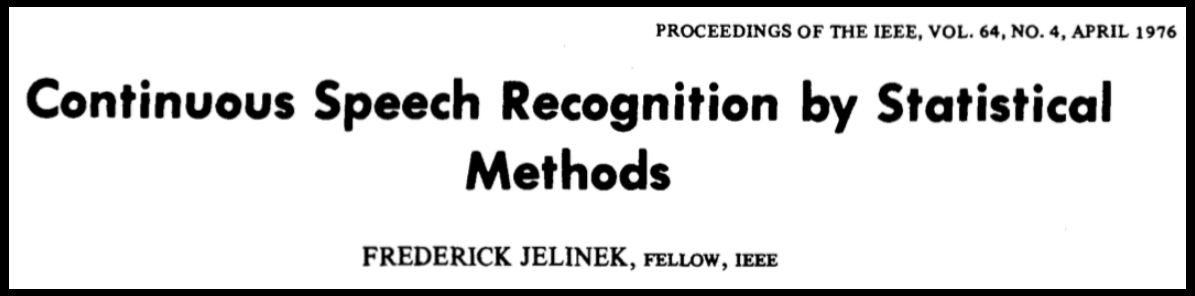


Word prediction



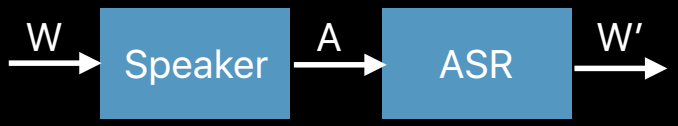
Deep Learning in Speech Recognition

Fundamental Equation of Speech Recognition


PROCEEDINGS OF THE IEEE, VOL. 64, NO. 4, APRIL 1976

Continuous Speech Recognition by Statistical Methods

FREDERICK JELINEK, FELLOW, IEEE



$$\hat{W} = \underset{W}{\operatorname{argmax}}\, p(W|A) = \underset{W}{\operatorname{argmax}}\, p(A|W)p(W) = \underset{W}{\operatorname{argmax}}\{\ln p(A|W) + \ln p(W)\}$$

$$\hat{W} = \underset{W}{\operatorname{argmax}}\{\lambda \ln p(A|W) + \ln p(W)\}$$

Acoustic Model

Language Model

Language Model



IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. PAMI-5, NO. 2, MARCH 1983 179

A Maximum Likelihood Approach to Continuous Speech Recognition

LALIT R. BAHL, MEMBER, IEEE, FREDERICK JELINEK, FELLOW, IEEE, AND ROBERT L. MERCER

$$p(W) = p(w_1 w_2 w_3 w_4 w_5) = p(w_1)p(w_2|w_1)p(w_3|w_1 w_2)p(w_4|w_1 w_2 w_3)p(w_5|w_1 w_2 w_3 w_4)$$

$$p(W) \cong p(w_1)p(w_2|w_1)p(w_3|w_1 w_2)p(w_4|w_2 w_3)p(w_5|w_3 w_4)$$

Trigram

$$p(w_i|w_{i-1}) = \frac{c(w_{i-1} w_i)}{\sum_{w_i} c(w_{i-1} w_i)}$$

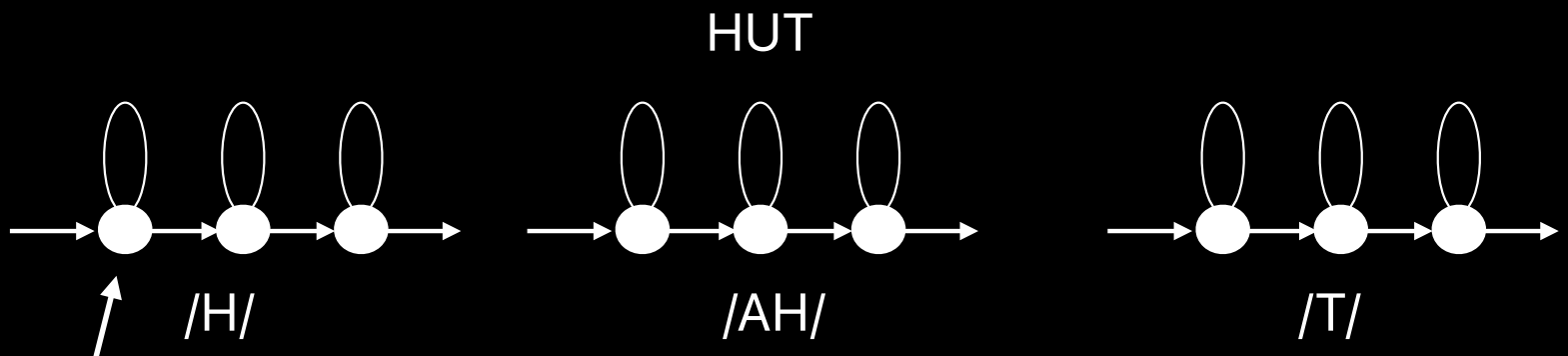
ML estimate

$$r^* = (r+1)\frac{n_{r+1}}{n_r}$$

Good Turing Smoothing

n_r is the number of n-grams seen exactly r times

Acoustic Model (Hidden Markov Models)



$$p(a_t|s_j) = \sum_{i=1}^{I} \alpha_i \mathcal{N}(a_t, \mu_{ij}, \Sigma_{ij})$$

24 IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-23, NO. 1, FEBRUARY 1975

The DRAGON System—An Overview

JAMES K. BAKER

PROCEEDINGS OF THE IEEE, VOL. 64, NO. 4, APRIL 1976

Continuous Speech Recognition by Statistical Methods

FREDERICK JELINEK, FELLOW, IEEE

Neural Networks for Speech Recognition in the 1990s

328 IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Merging Multilayer Perceptrons and Hidden Markov Models: Some Experiments in Continuous Speech Recognition

H. Bourlard[1], N. Morgan[2]

TR-89-033

July 1989

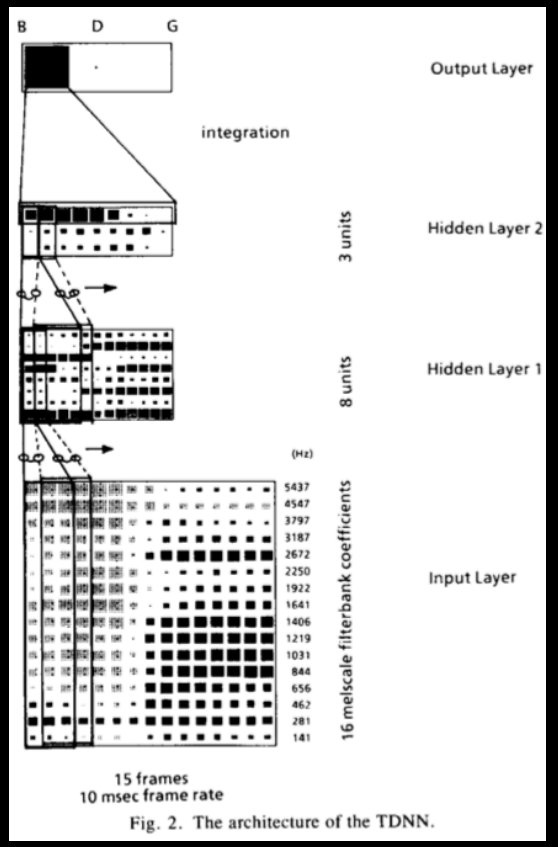
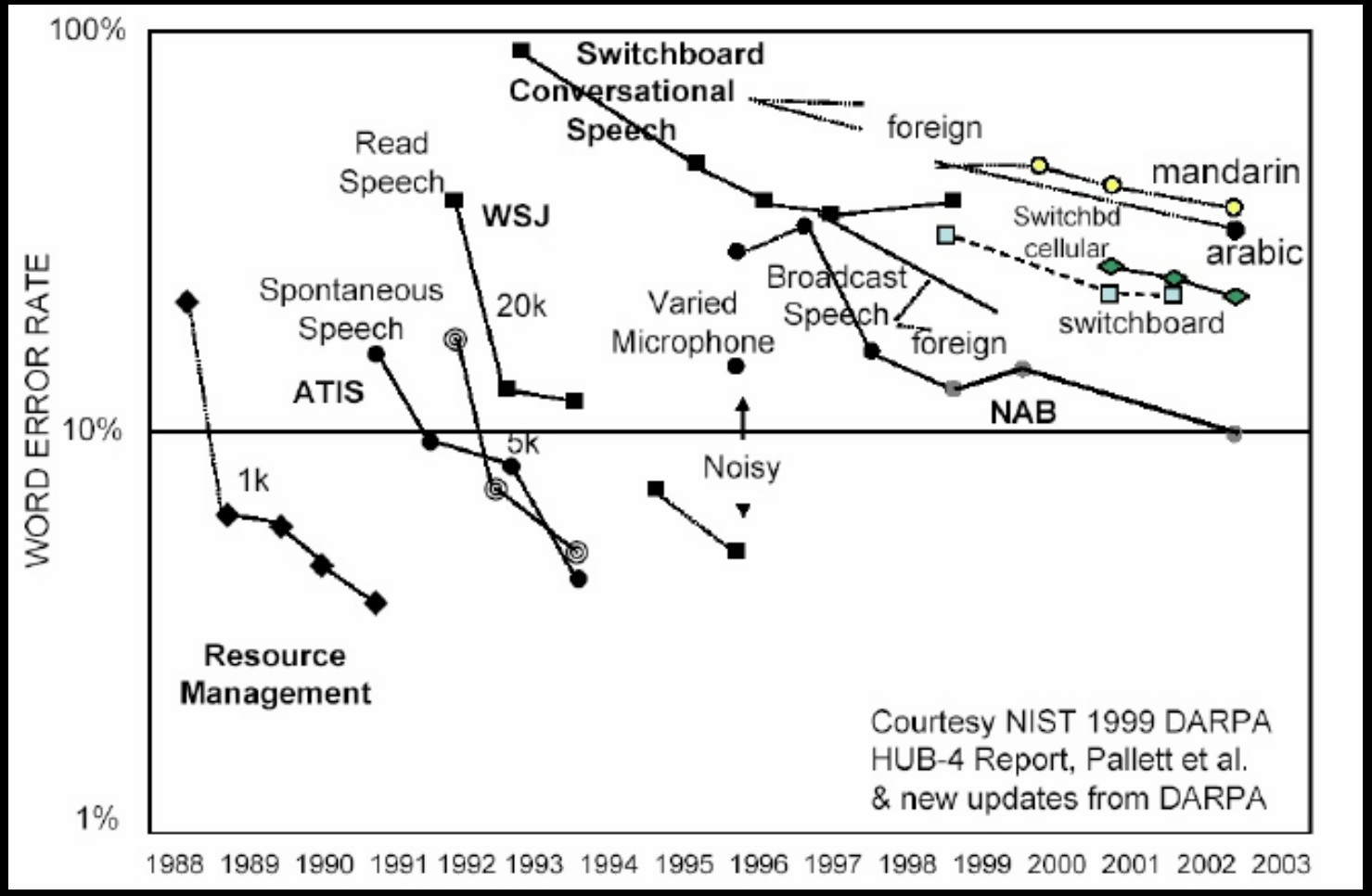


Fig. 2. The architecture of the TDNN.

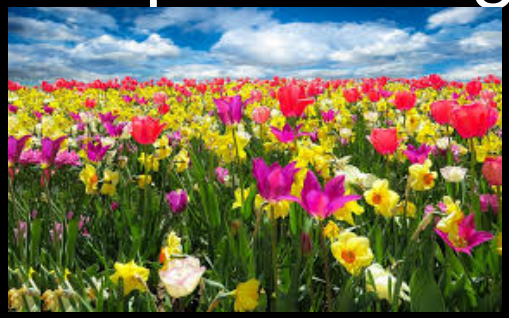
Neural Network Winter for Speech Recogntion



Open Challenge Tasks (DARPA)



Deep Learning



Reducing the Dimensionality of Data with Neural Networks

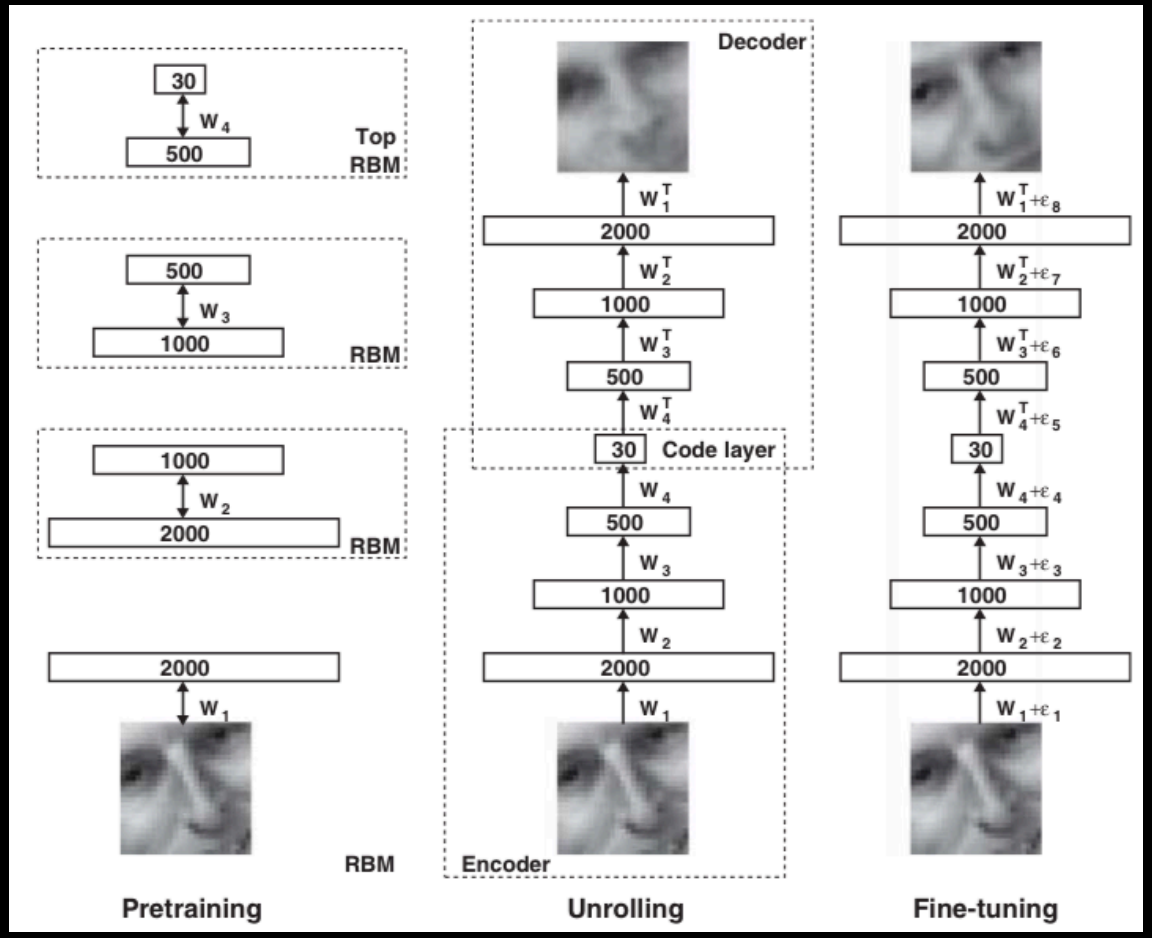
G. E. Hinton* and R. R. Salakhutdinov

28 JULY 2006 VOL 313 SCIENCE www.sciencemag.org

Deep Boltzmann Machines

Ruslan Salakhutdinov
Department of Computer Science
University of Toronto
rsalakhu@cs.toronto.edu

Geoffrey Hinton
Department of Computer Science
University of Toronto
hinton@cs.toronto.edu



Deep Belief Networks => Deep Neural Networks

INTERSPEECH 2010

Investigation of Full-Sequence Training of Deep Belief Networks for Speech Recognition

Abdel-rahman Mohamed [1*], *Dong Yu*[2], *Li Deng*[2]

[1] Department of Computer Science, University of Toronto, Toronto, ON Canada
[2] Microsoft Research, Redmond, WA USA

INTERSPEECH 2011

Conversational Speech Transcription Using Context-Dependent Deep Neural Networks

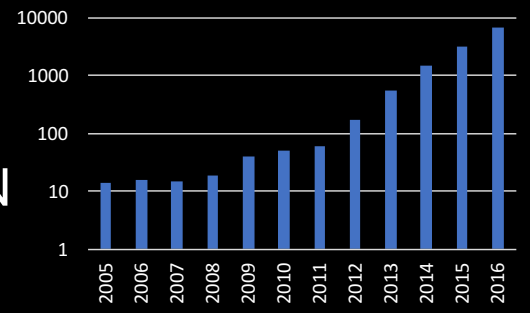
Frank Seide[1], *Gang Li,*[1] *and Dong Yu*[2]

[1]Microsoft Research Asia, Beijing, P.R.C.
[2]Microsoft Research, Redmond, USA

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

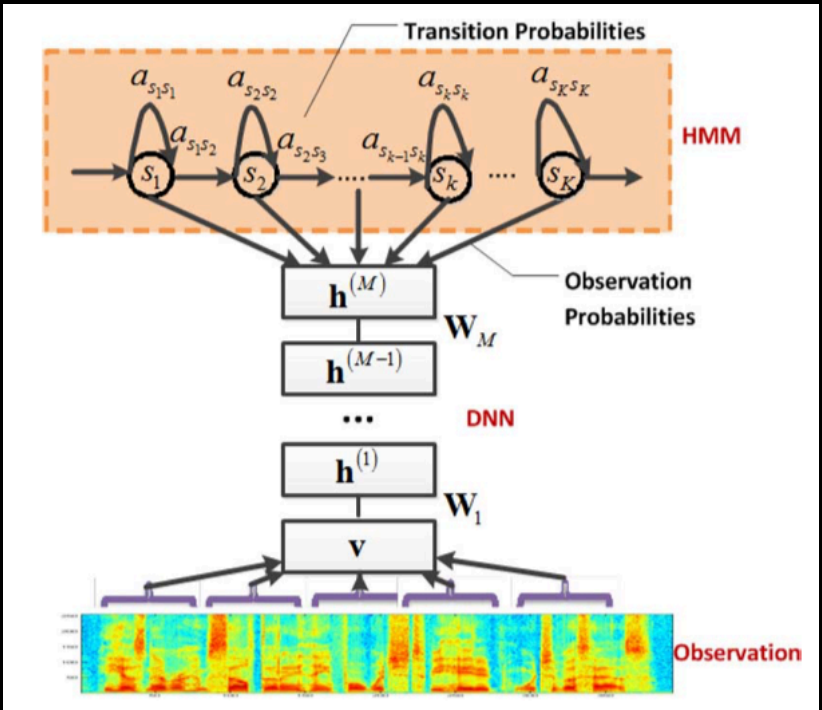
George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*

#Papers with DNN



Deep Learning for Speech

(Deng et al., 2010)



DNNs for large vocabulary

800 input features
5 layer network
1000 neurons per layer
8000 output labels
12 Million weights

Training

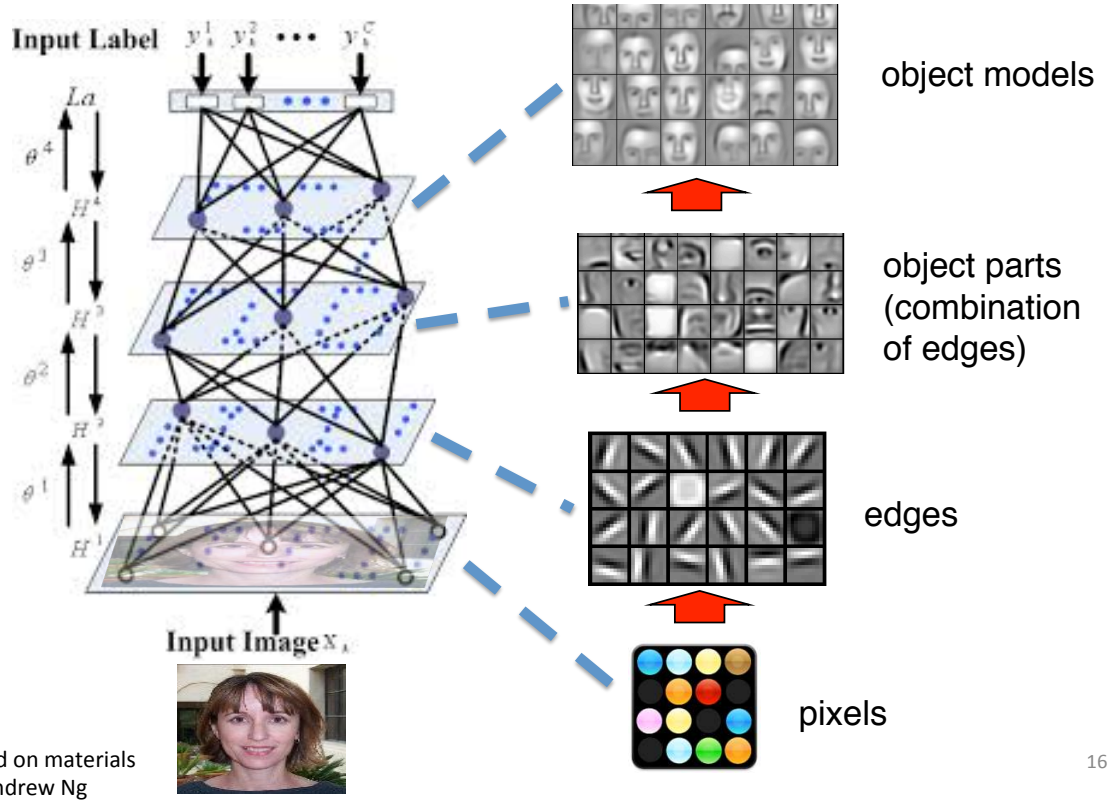
1000 hours of speech with transcripts
1 week training time on GPU farm

Deep Neural Networks: What was new?

- Algorithms:
  - Direct modeling of context-dependent (tied triphone states) through the DNN
  - ~~Unsupervised Pretraining~~
  - Deeper networks
- Data: Large amounts of data
- Compute: GPUs

DNN on Face Images (2012)

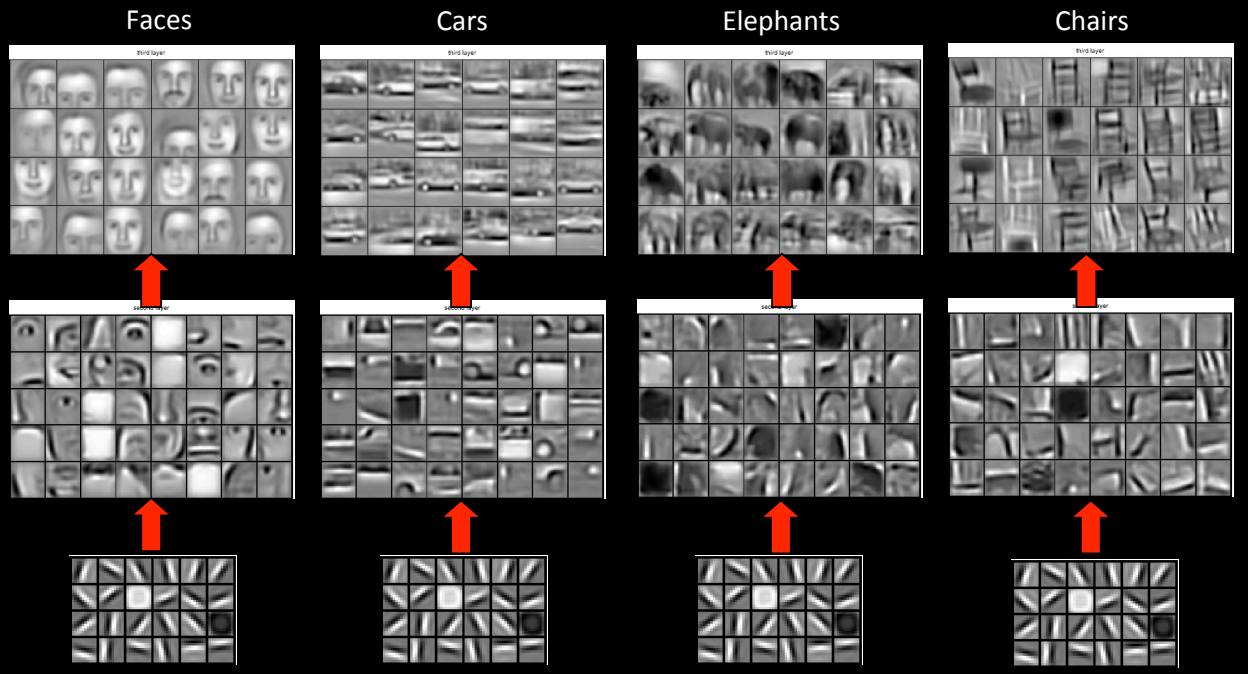
Deep Belief Net on Face Images



Based on materials
by Andrew Ng

ImageNet neurons (2012)

Examples of learned object parts from object categories

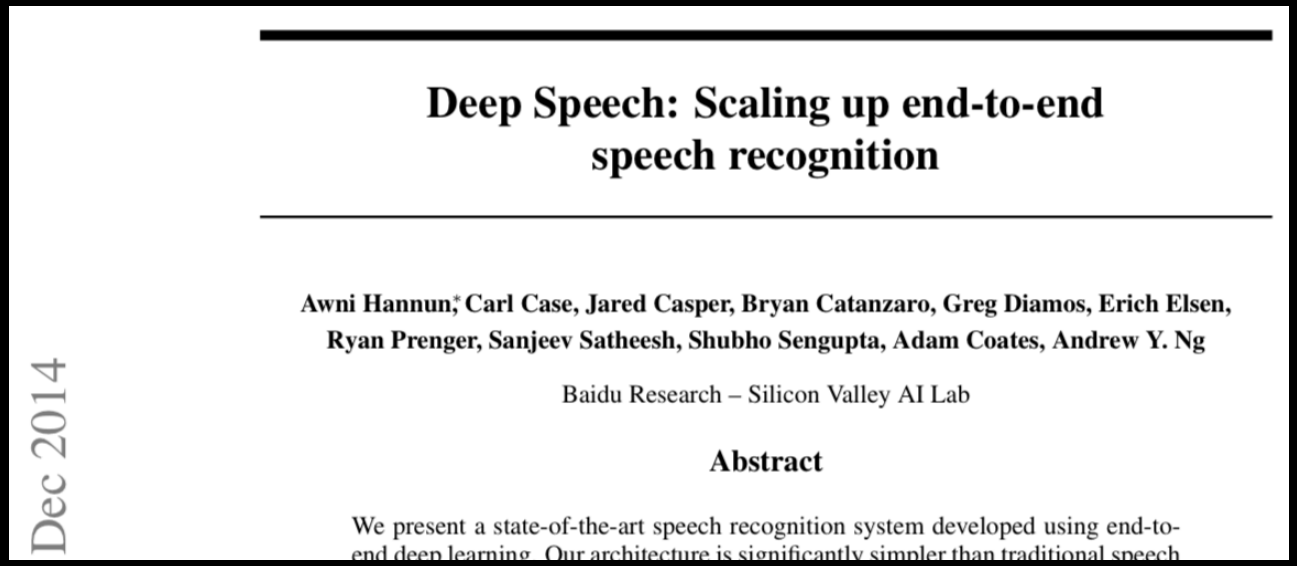


Credit Andrew Ng

Deep Learning in Speech Recognition

- HMMs: yeap, still there
- Training criterion: Cross-entropy => MMI Sequence-level
- Features: Mel-Frequency Cepstral Coefficients (MFCC)=> FilterBanks
- Training & regularizetion: batch normalization, distributed SGD, dropout, etc
- Acoustic Modeling: CNN/CTC/CLDNN
- Language Modeling: RNNs (in rescoring pass)
- Data: even more data, and more diverse (noisy, accents, etc)
- Compute: even more GPUs

End-to-end Learning



Dec 2014

Deep Speech: Scaling up end-to-end speech recognition

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng

Baidu Research – Silicon Valley AI Lab

Abstract

We present a state-of-the-art speech recognition system developed using end-to-end deep learning. Our architecture is significantly simpler than traditional speech

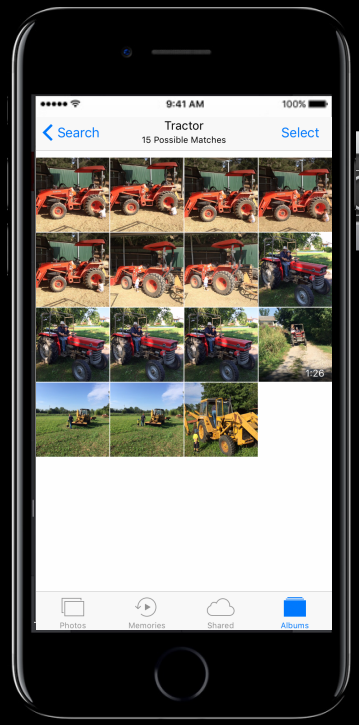
- Language Modeling uses 1000 more training data than Acoustic modeling
- Dictionaries can help with never seen words

Still lots to do ...

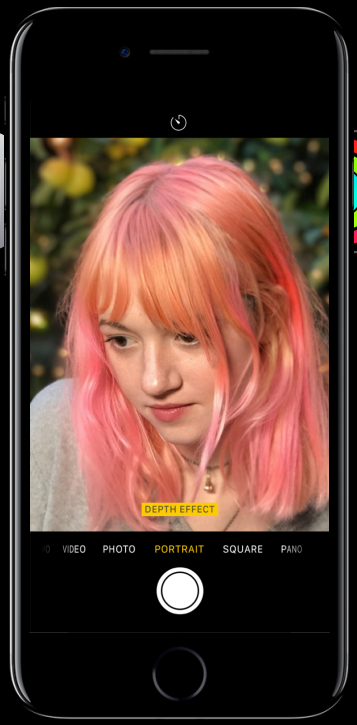


Machine Learning across Apple Products

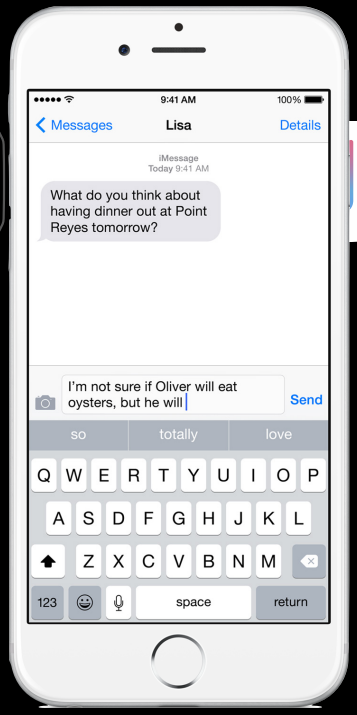
On-device scene recognition



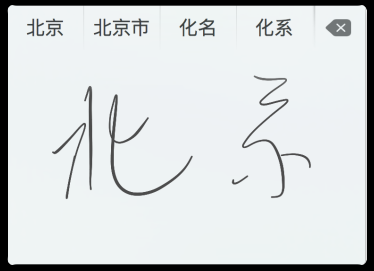
Portrait Mode



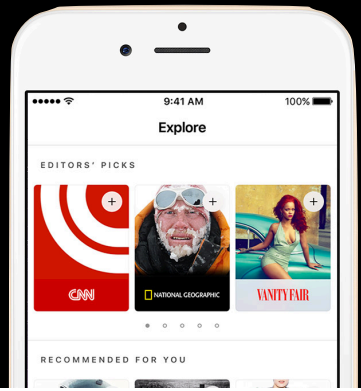
Language modeling



Handwriting recognition



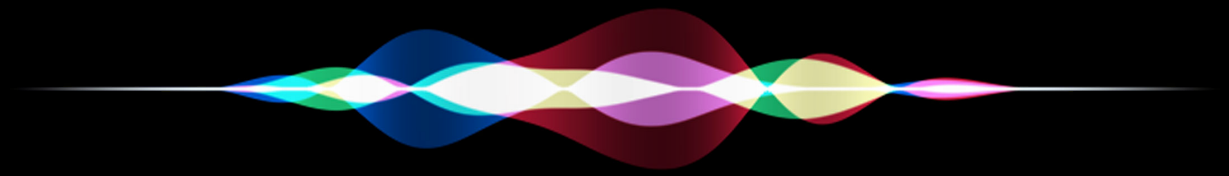
News recommendation



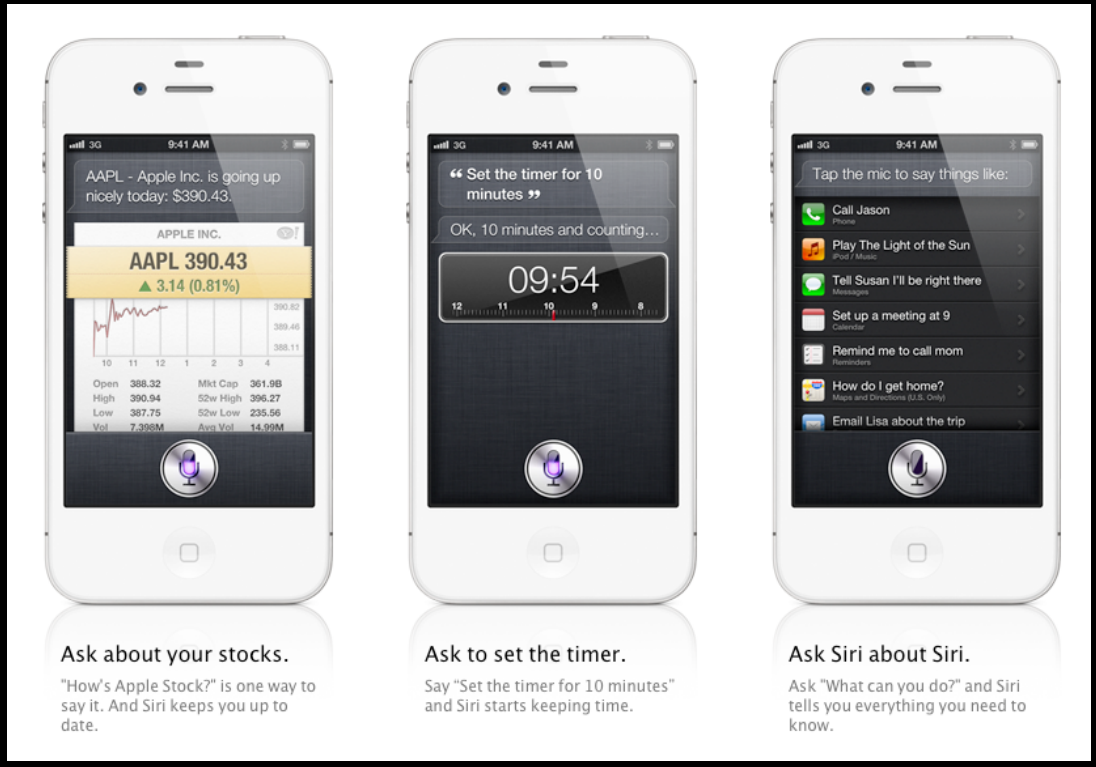
Intelligent assistant



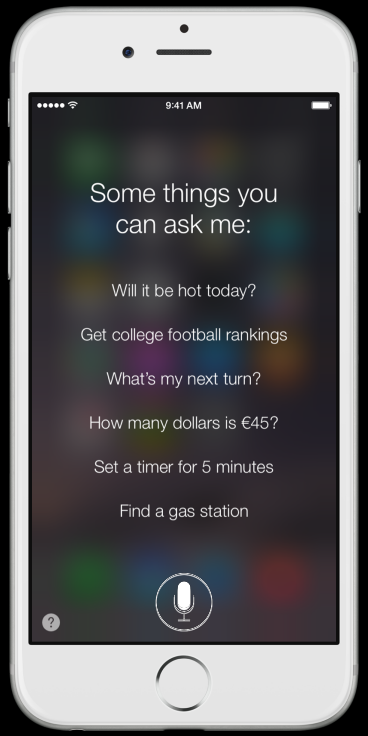
Introduction to Siri



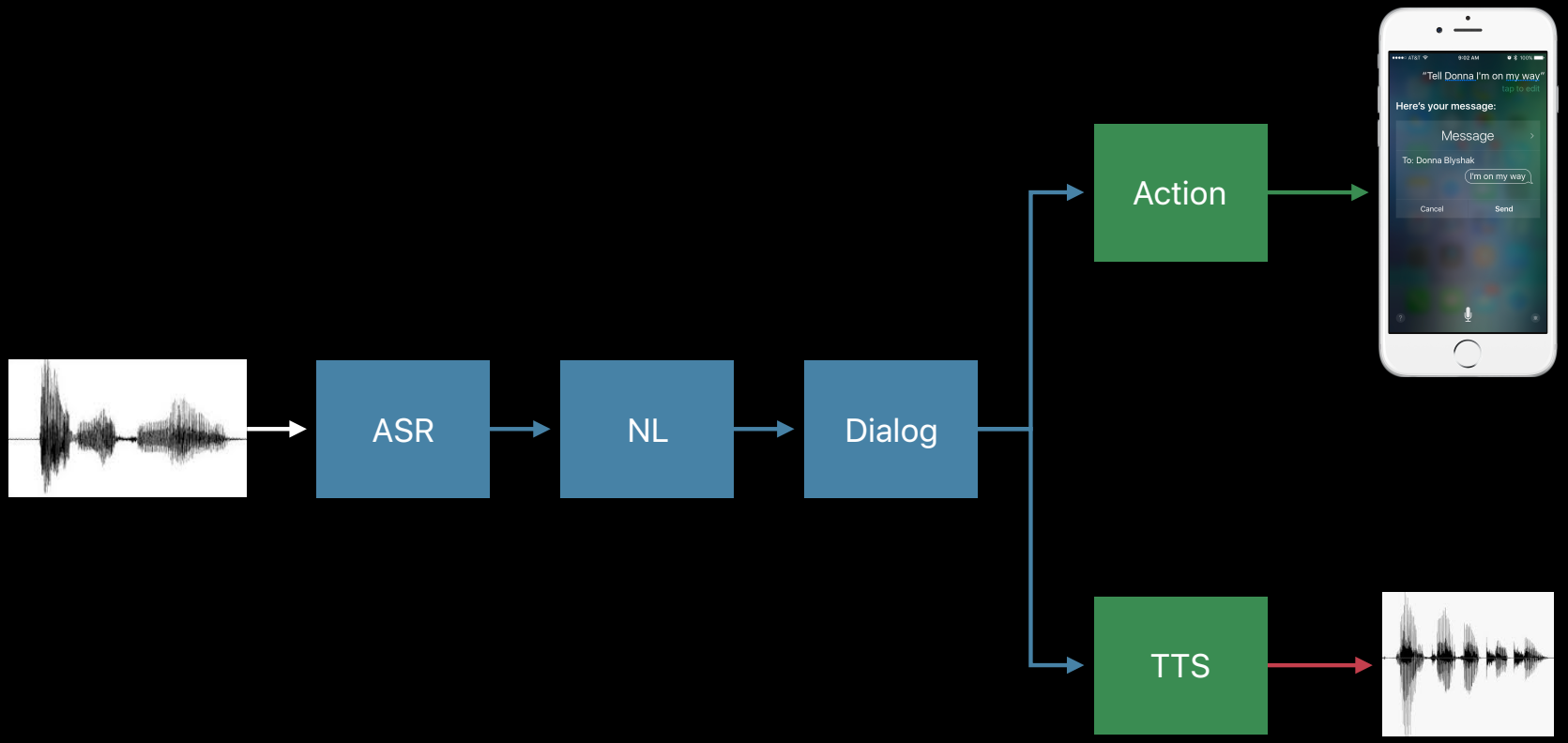
Siri (2011)


AAPL - Apple Inc. is going up nicely today: $390.43.
APPLE INC.
AAPL 390.43
▲ 3.14 (0.81%)
Open 388.32
High 390.94
Low 387.75
Mkt Cap 361.9B
52w High 396.27
52w Low 235.56
Ask about your stocks.
"How's Apple Stock?" is one way to say it. And Siri keeps you up to date.
“Set the timer for 10 minutes”
OK, 10 minutes and counting…
09:54
Ask to set the timer.
Say “Set the timer for 10 minutes” and Siri starts keeping time.
Tap the mic to say things like:
Call Jason
Play The Light of the Sun
Tell Susan I'll be right there
Set up a meeting at 9
Remind me to call mom
How do I get home?
Email Lisa about the trip
Ask Siri about Siri.
Ask "What can you do?" and Siri tells you everything you need to know.



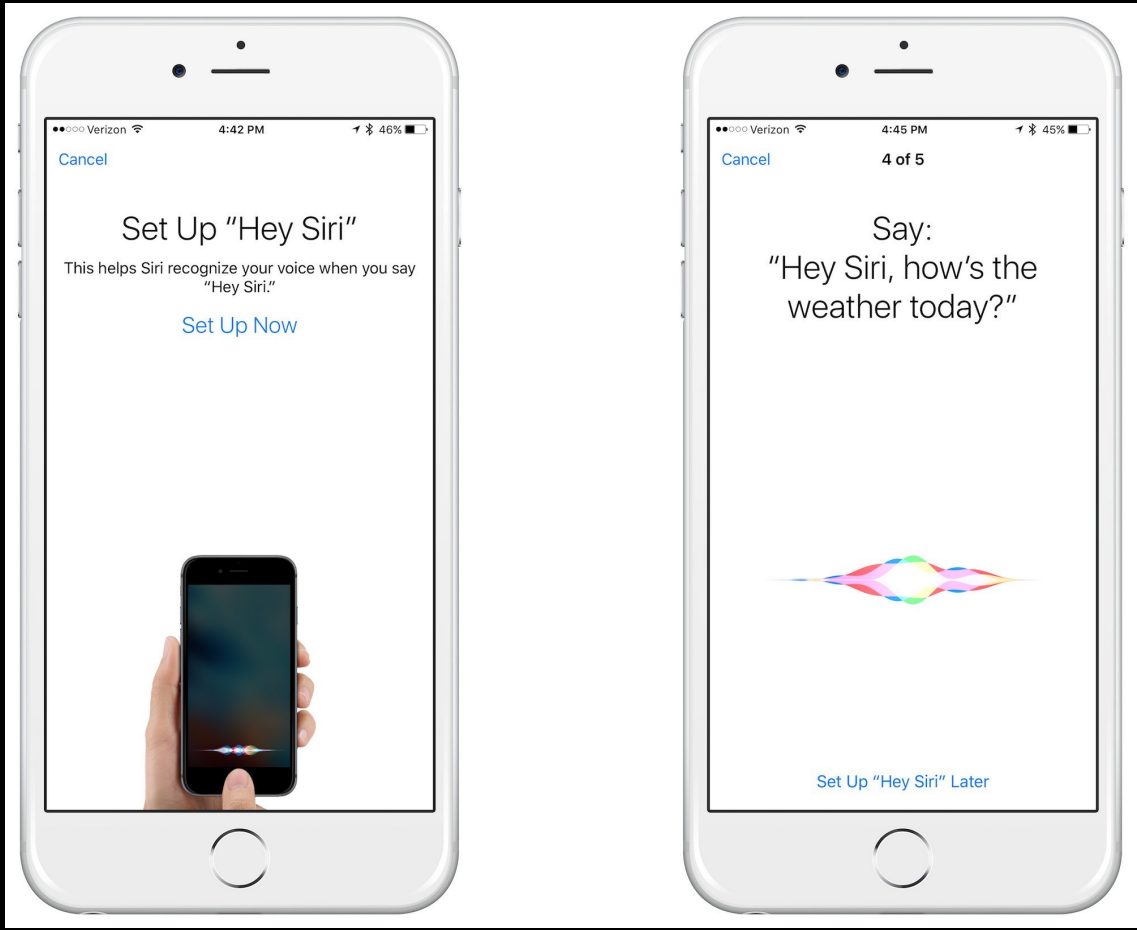

9:41 AM
Some things you can ask me:
Will it be hot today?
Get college football rankings
What's my next turn?
How many dollars is €45?
Set a timer for 5 minutes
Find a gas station

Siri Architecture



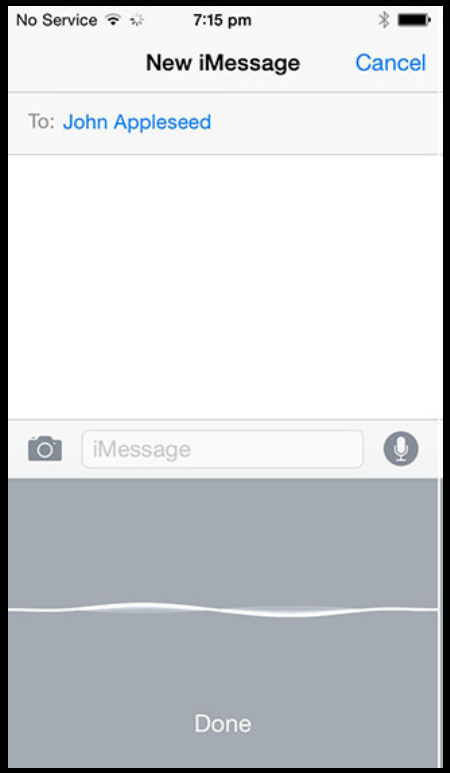

Monday
23
280
W

Hands-Free Siri

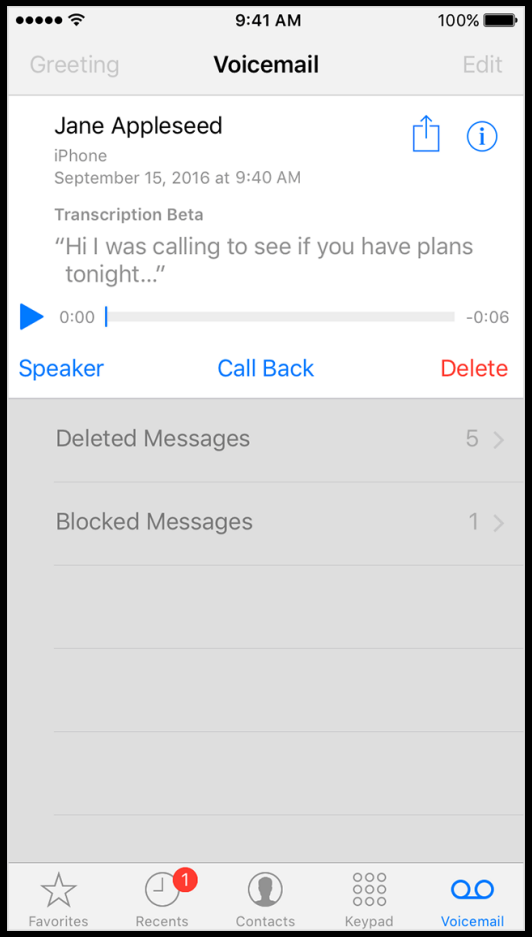


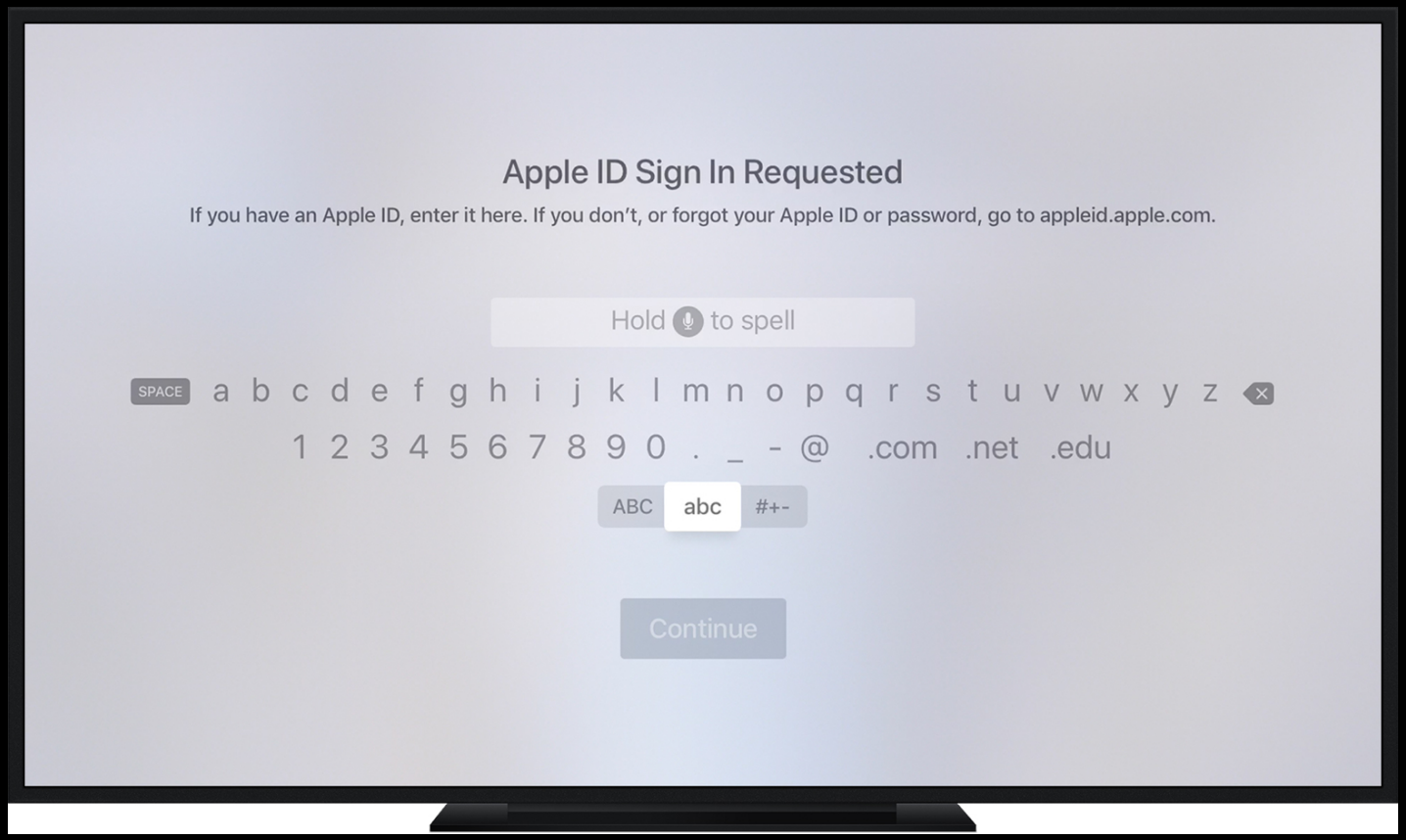
Dictation


Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
123
space
return


No Service
7:15 pm
New iMessage
Cancel
To: John Appleseed
iMessage
Done

Voicemail transcription





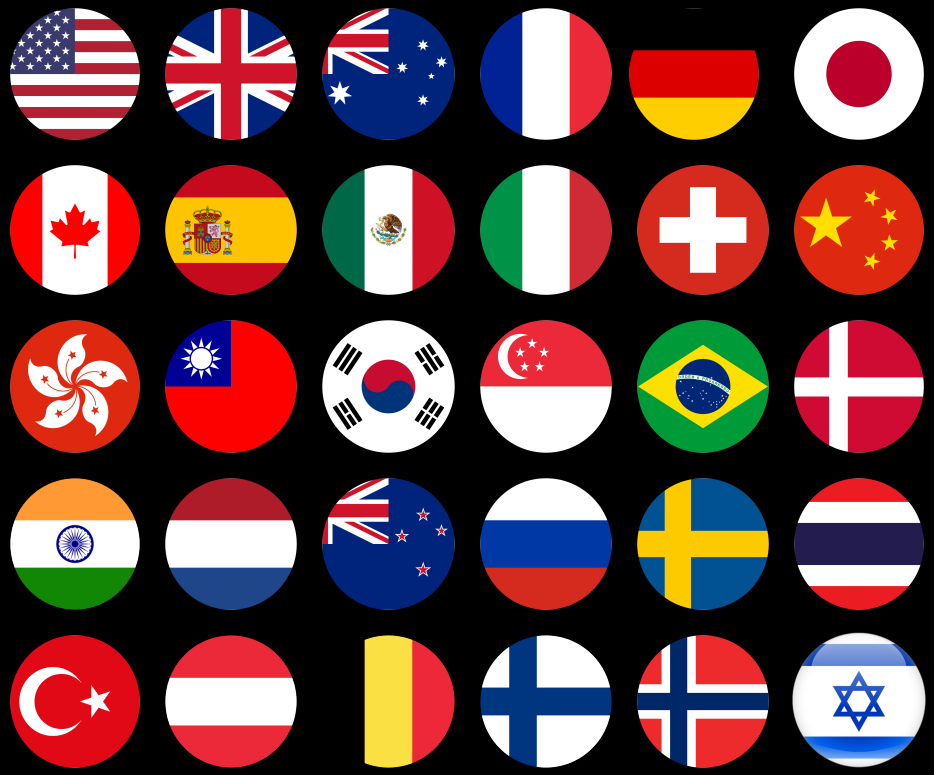
Entering password is very tedious with on-screen keyboard and trackpad

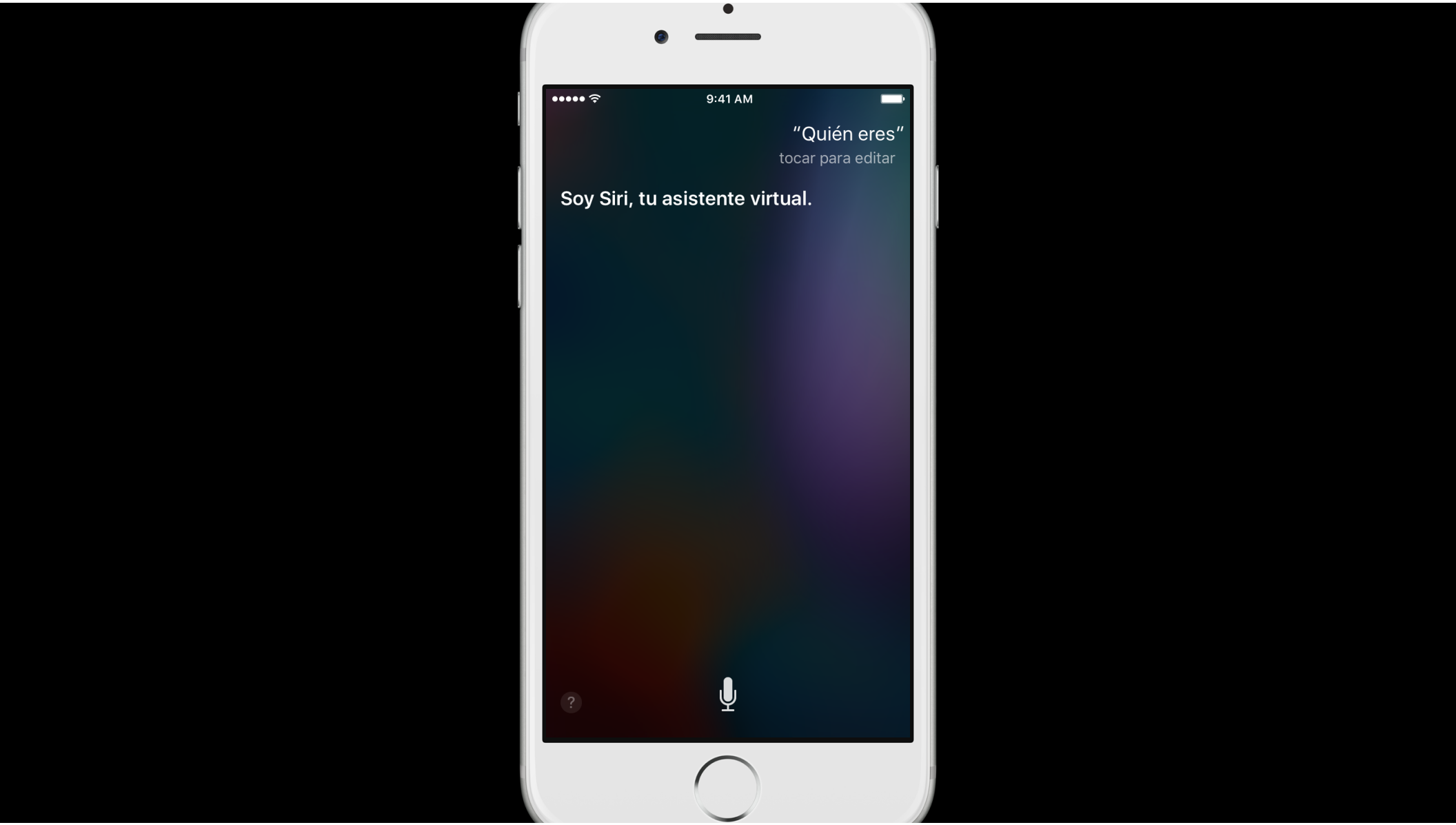
Apple does ***not*** want audio or logs of 3rd party passwords

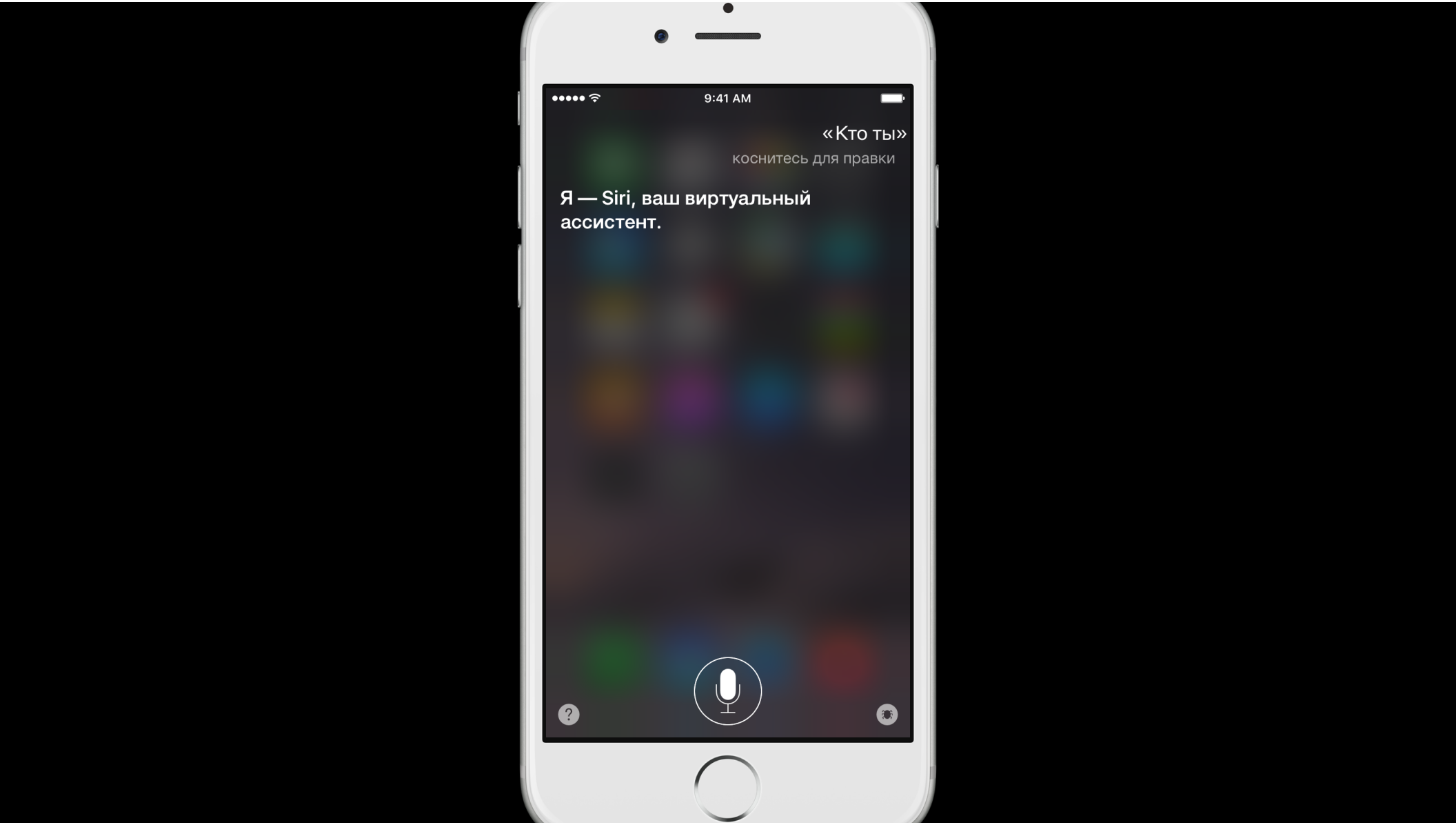
==> had to build solution running entirely on the ATV box

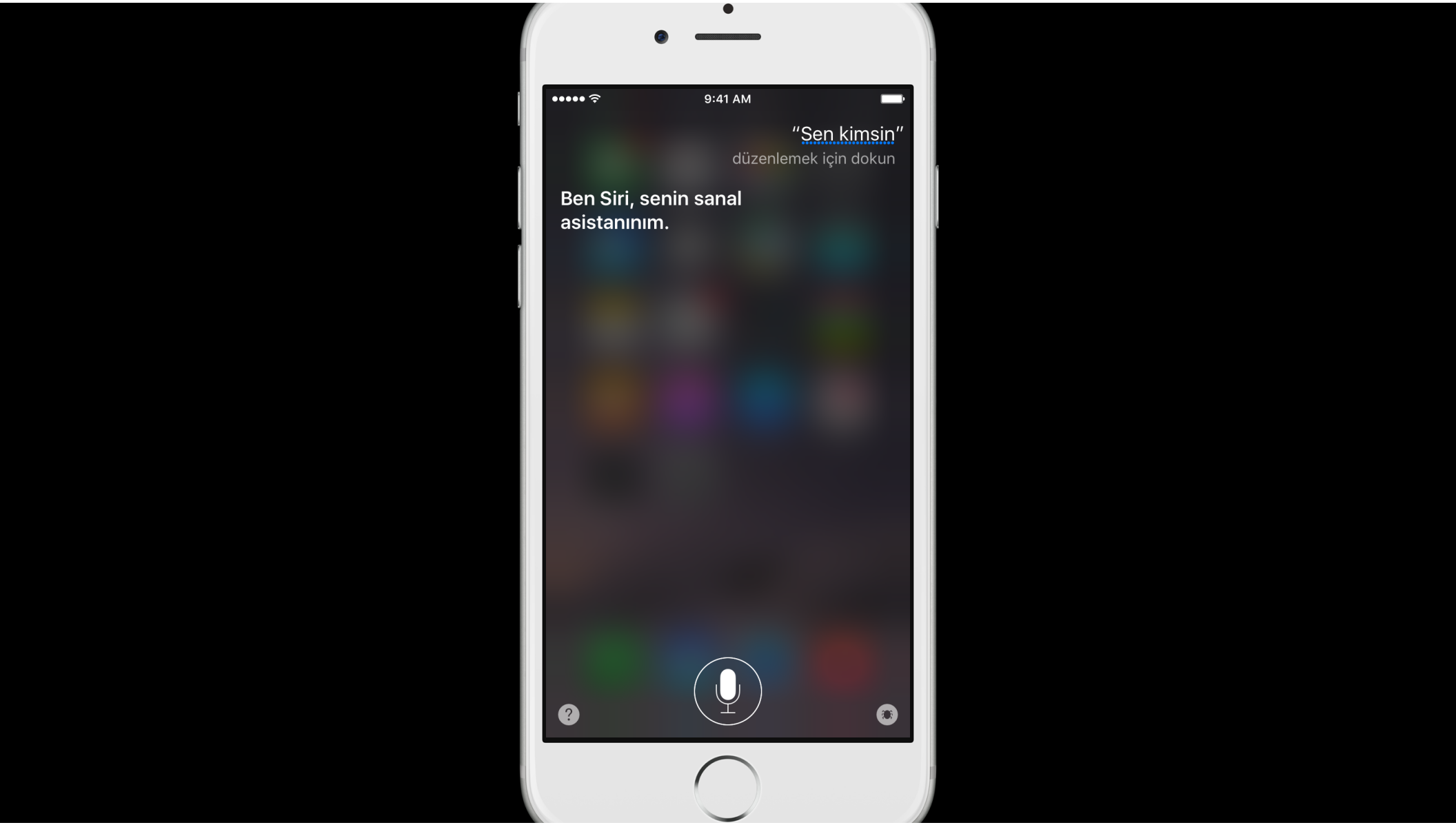
Apple Confidential–Internal Use Only

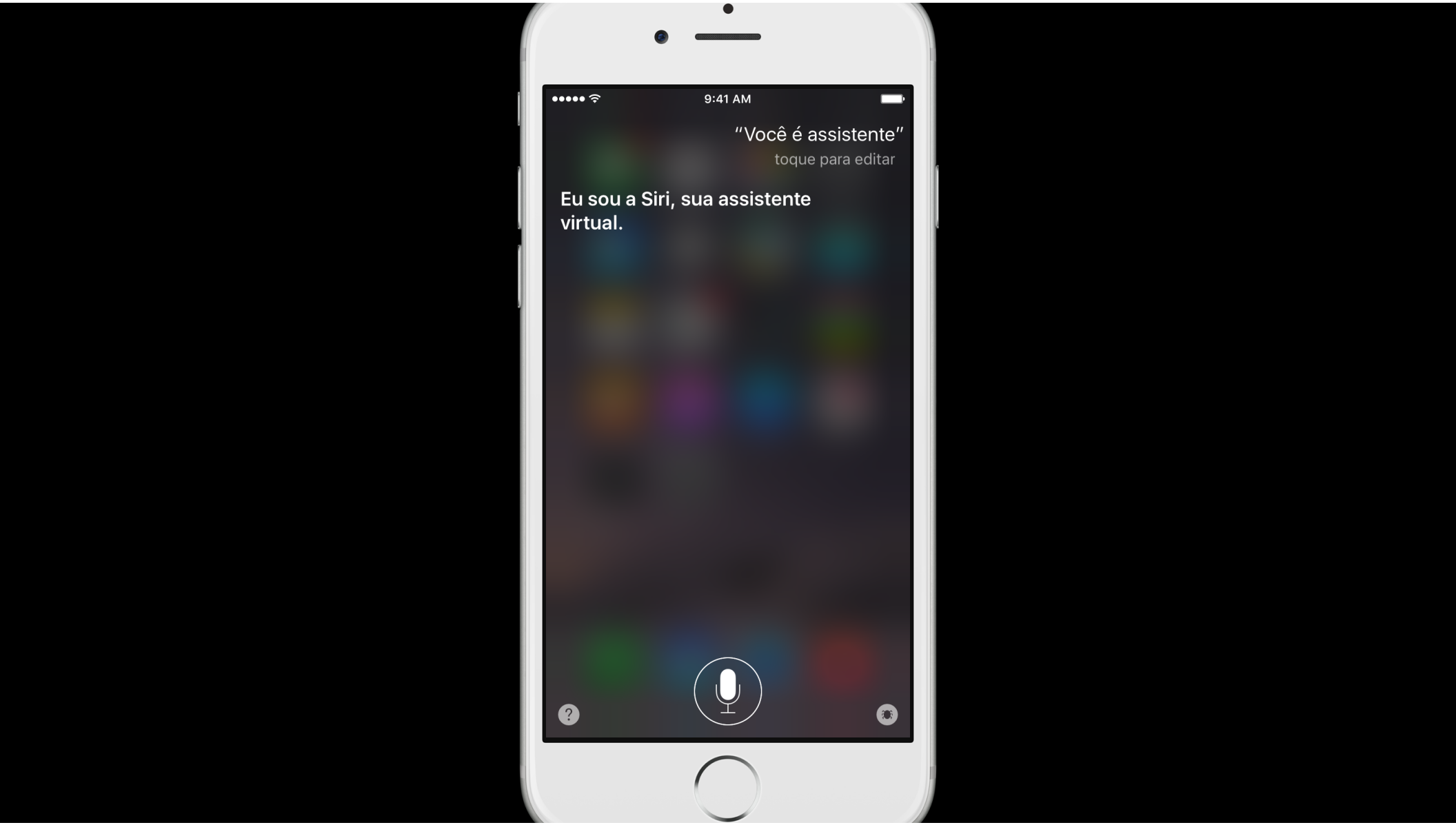
Global Reach

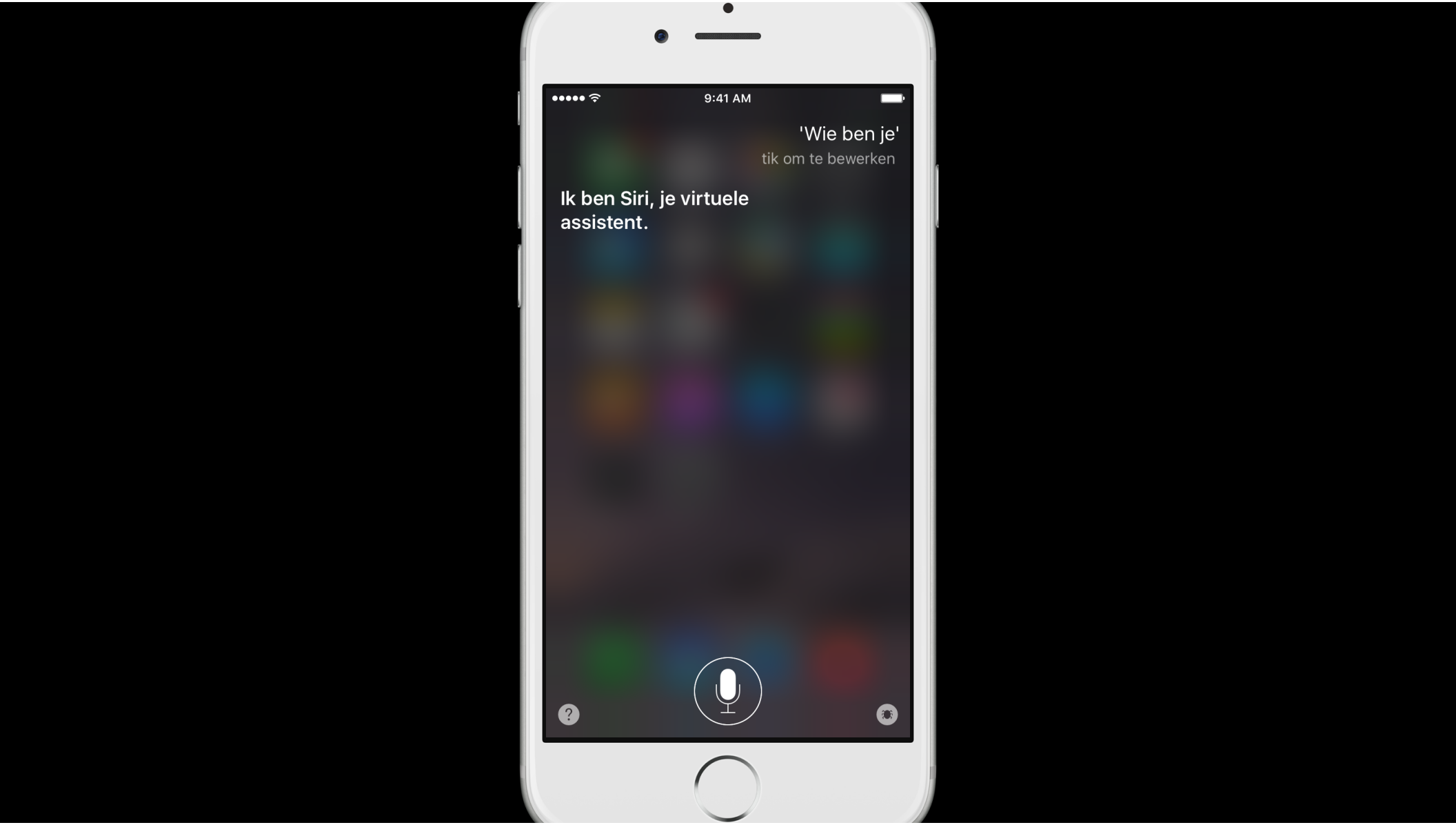


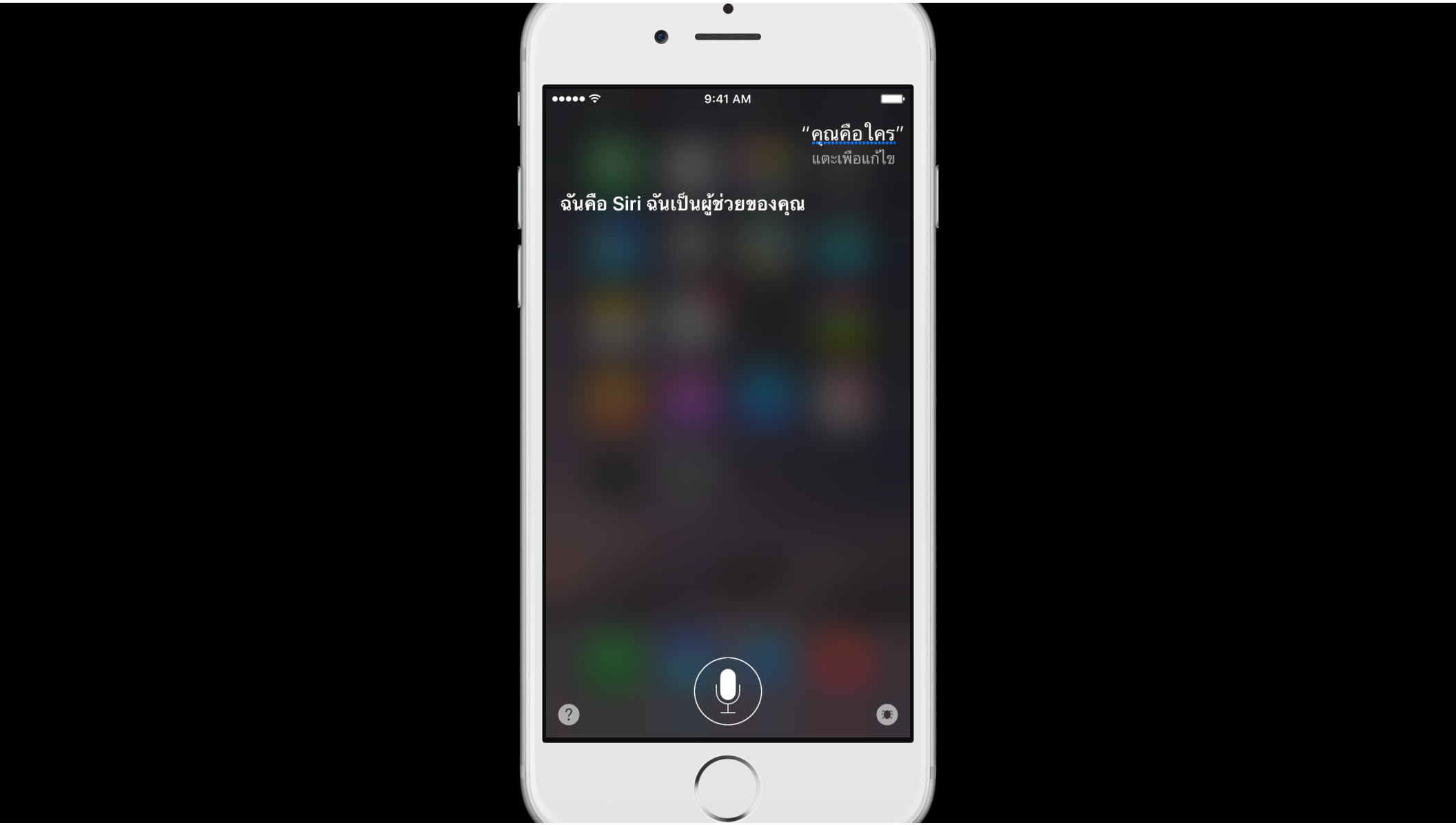

9:41 AM
"Quién eres"
tocar para editar
Soy Siri, tu asistente virtual.


9:41 AM
«Кто ты»
коснитесь для правки
Я — Siri, ваш виртуальный ассистент.

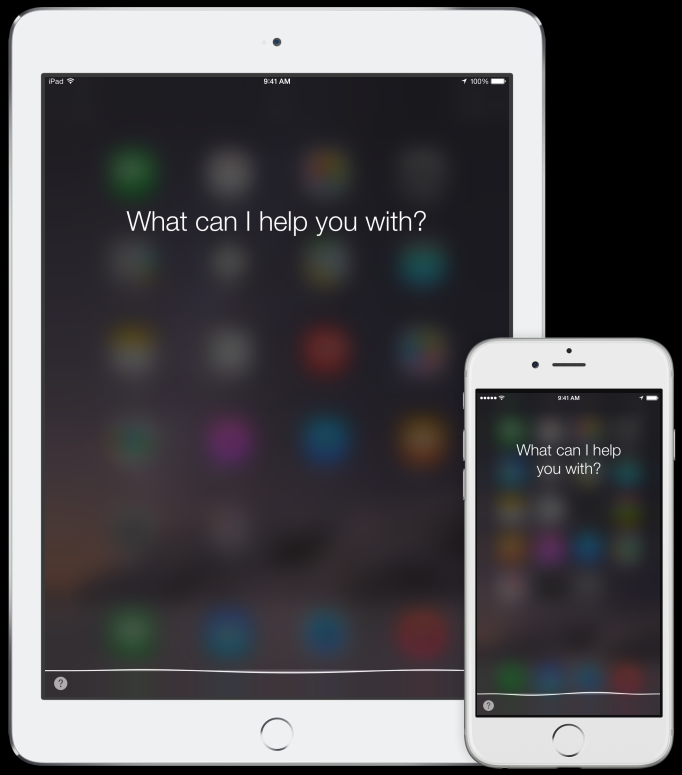

9:41 AM
"Sen kimsin"
düzenlemek için dokun
Ben Siri, senin sanal asistanınım.


9:41 AM
"Você é assistente"
toque para editar
Eu sou a Siri, sua assistente virtual.


9:41 AM
'Wie ben je'
tik om te bewerken
Ik ben Siri, je virtuele assistent.


9:41 AM
“คุณคือใคร”
แตะเพื่อแก้ไข
ฉันคือ Siri ฉันเป็นผู้ช่วยของคุณ

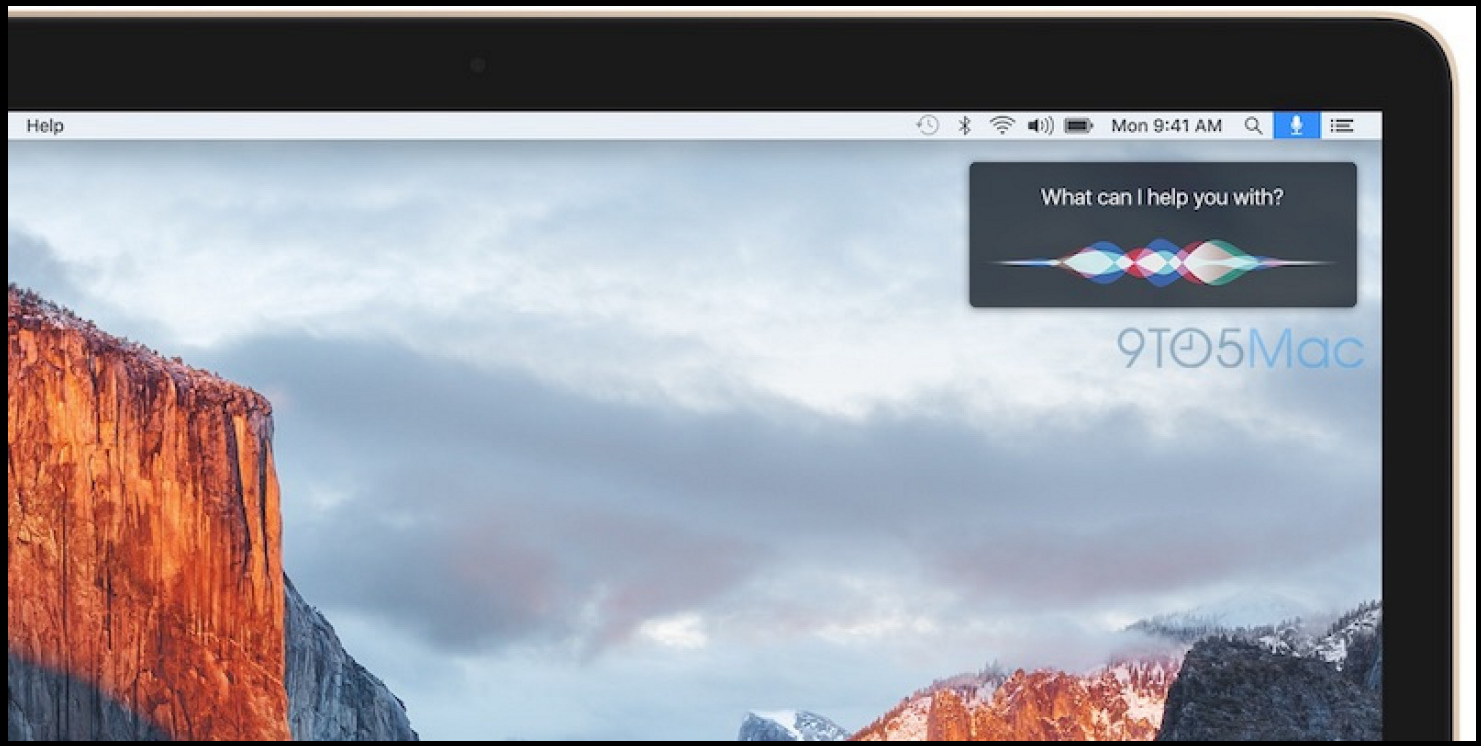
In many devices


iPad
9:41 AM
100%
What can I help you with?
9:41 AM
What can I help you with?


Hey Siri,



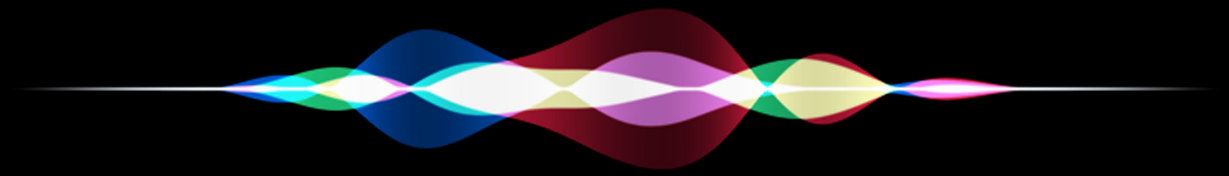
THE NATIONAL ACADEMY OF TELEVISION ARTS & SCIENCES


Help
Mon 9:41 AM
What can I help you with?
9TO5Mac

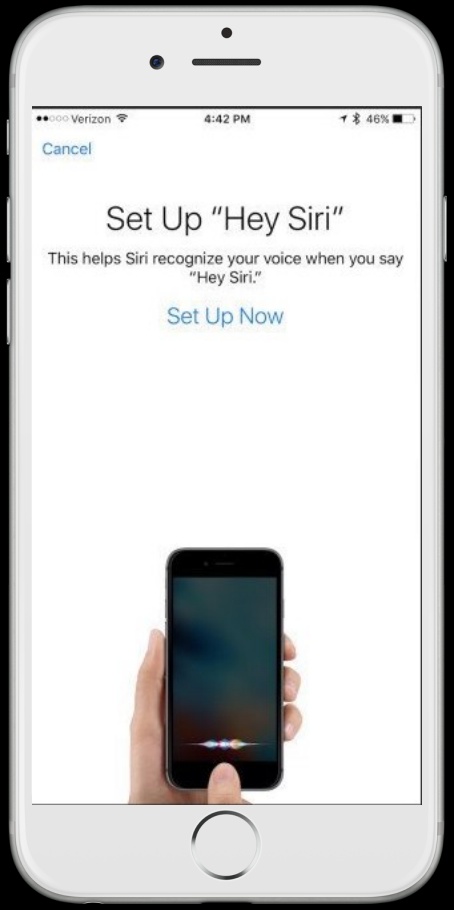




Deep Learning in Hey Siri

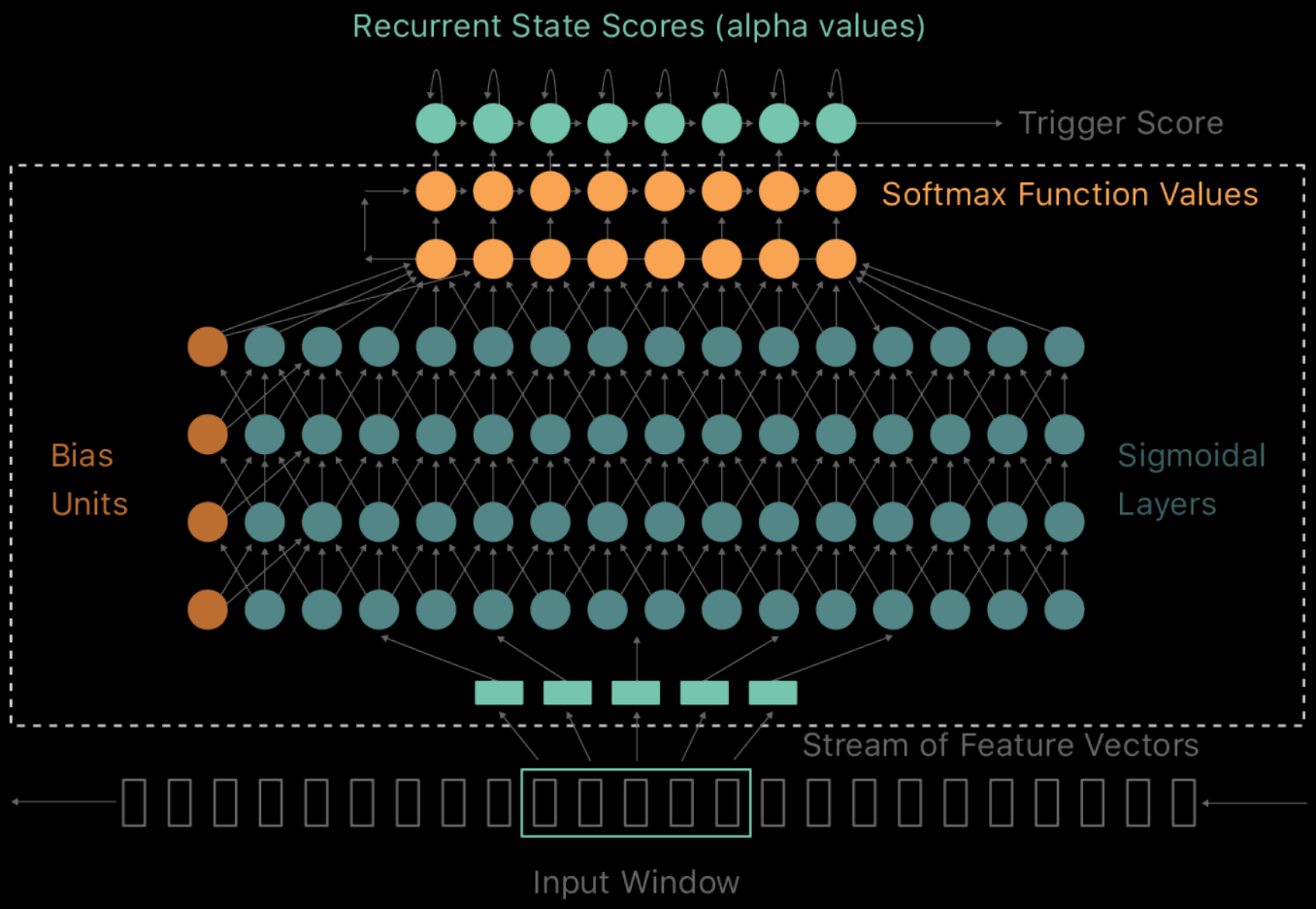


Hands-Free Siri

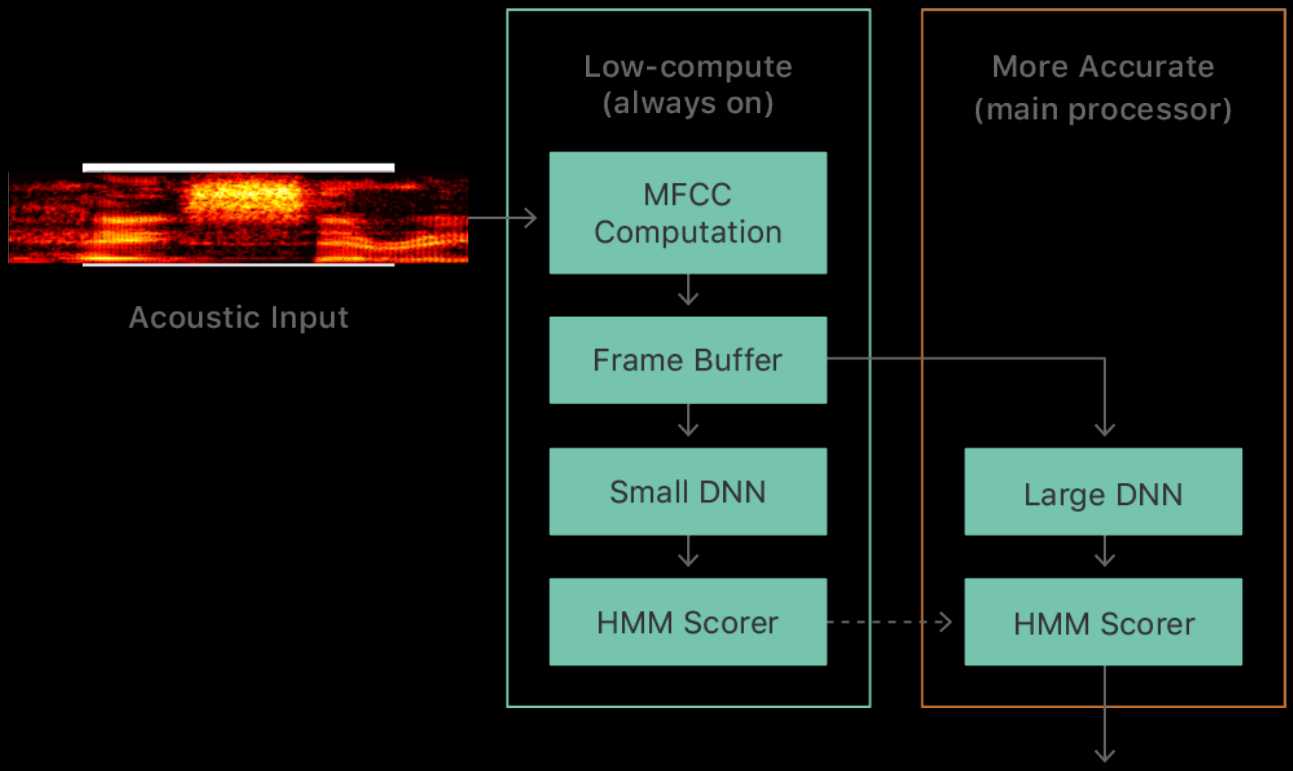




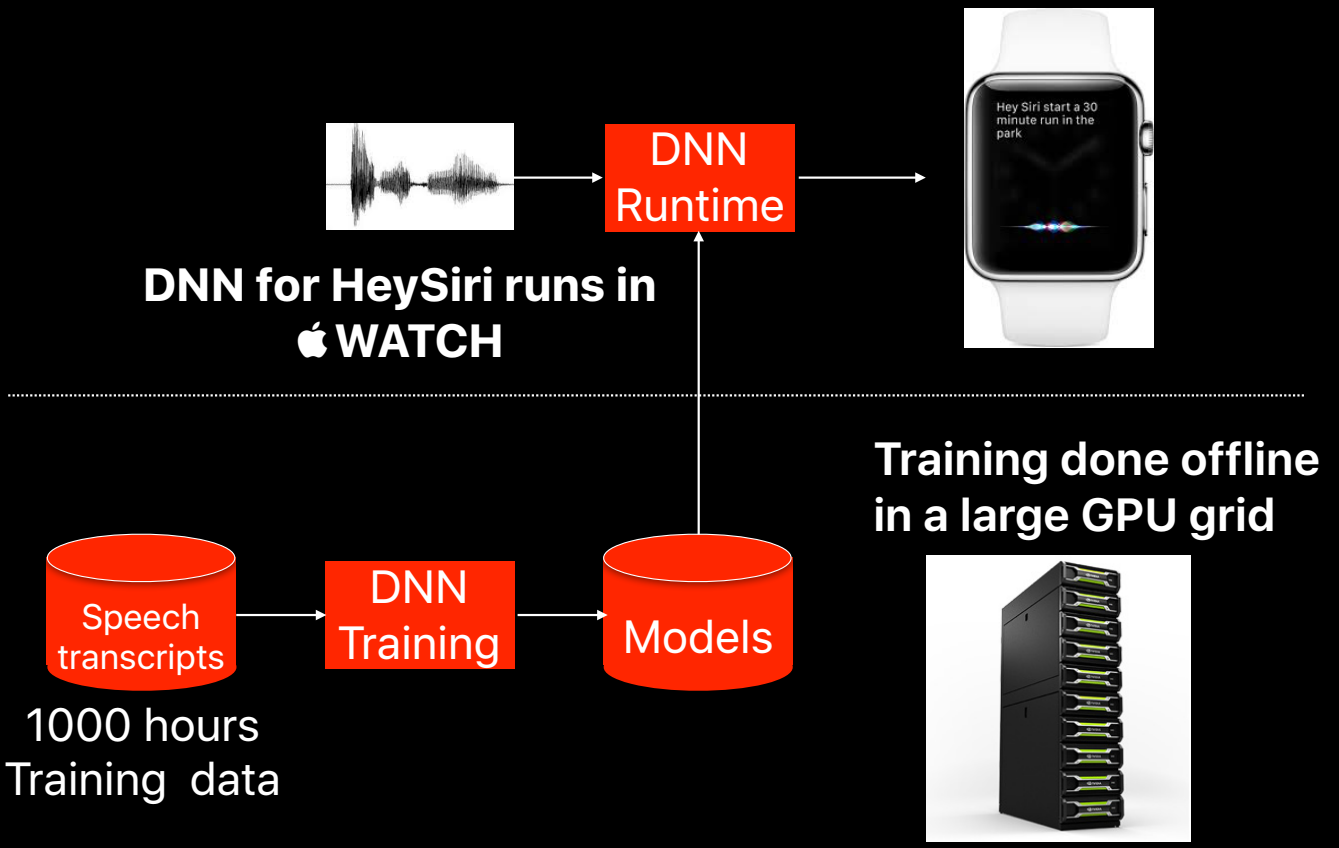
Hey Siri DNN


Recurrent State Scores (alpha values)
Trigger Score
Softmax Function Values
Bias
Units
Sigmoidal
Layers
Stream of Feature Vectors
Input Window

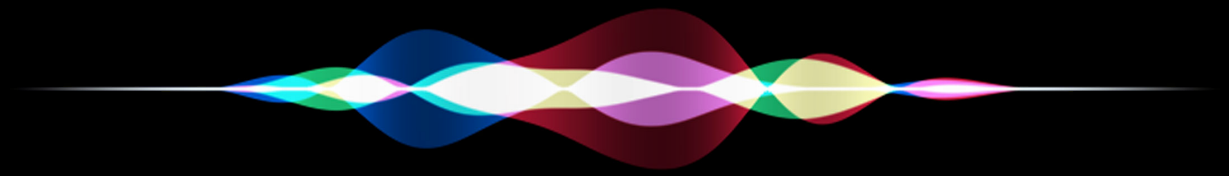
Two-pass Detection



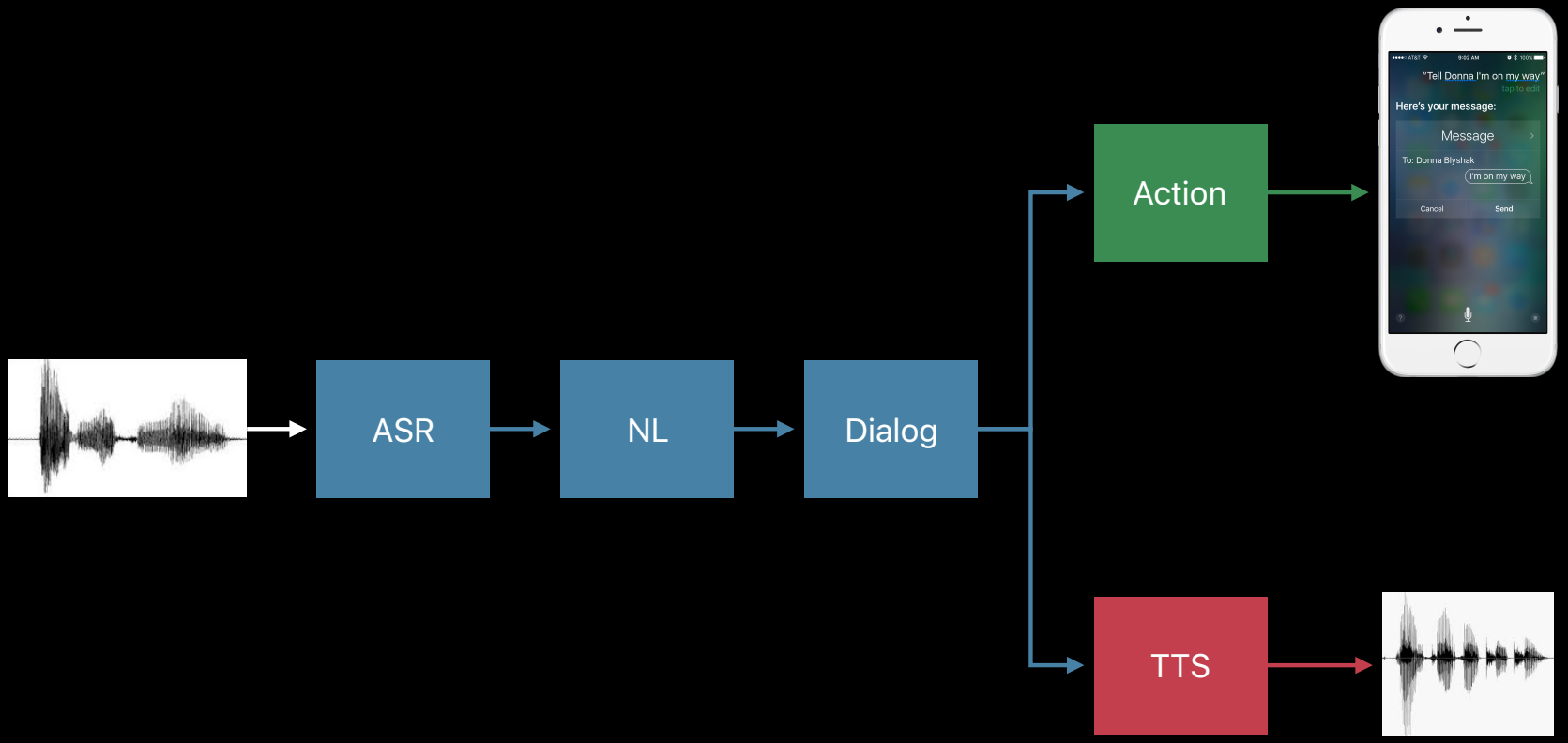
Computing for Deep Learning



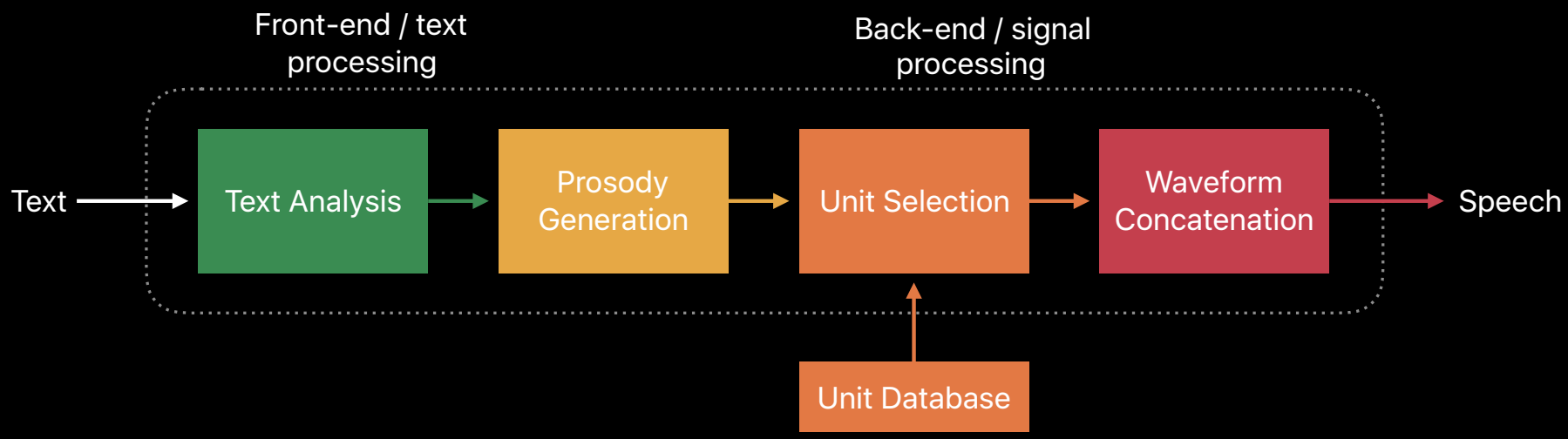
Deep Learning in Siri's TTS

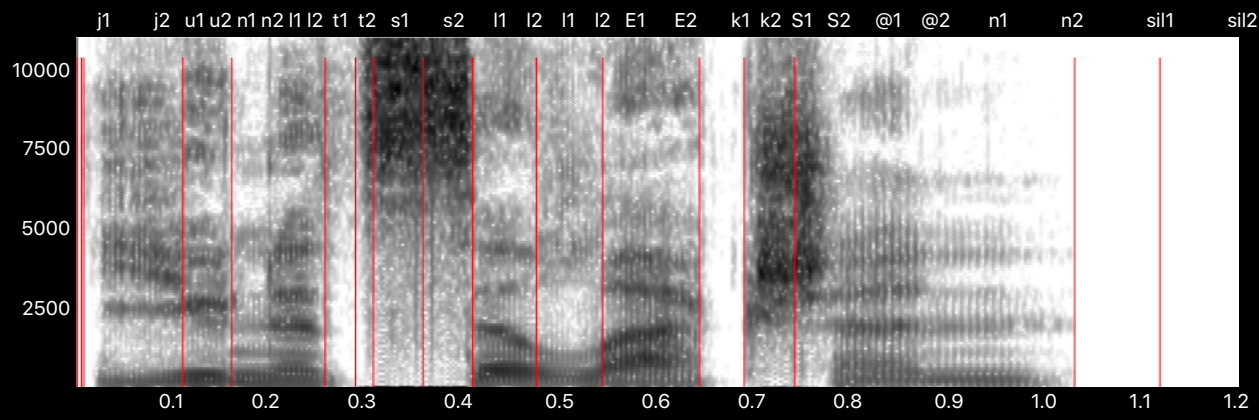


Siri Architecture



Concatenative TTS





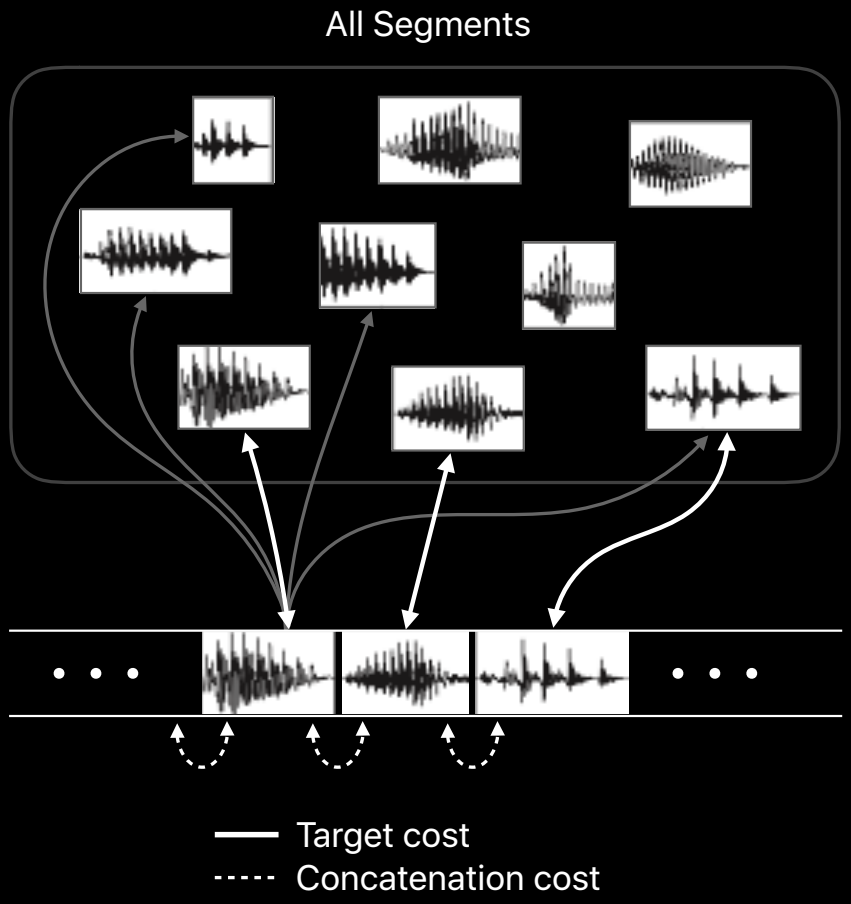
Database Segmentation



Final voice runs on iPhone using < 500MB:

- acoustic model
- unit index
- unit database (1-2 M half-phones)

Concatenative TTS



Viterbi search over all possible sequences of units to minimize total cost

$$C = w_t \sum_{n=1}^{N} T(u_n) + w_c \sum_{n=1}^{N-1} C(u_n, u_{n+1})$$

Target cost **Concatenation cost**

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," ICASSP, 1996.

Acoustic Features

prosodic	Log-F0 (target intonation) ΔF0 (how intonation evolves at concatenation point) Log-duration
spectral	MFCCs (target spectrum) ΔMFCCs (how spectrum evolves at concatenation point)

A total of 58 means and 58 variances

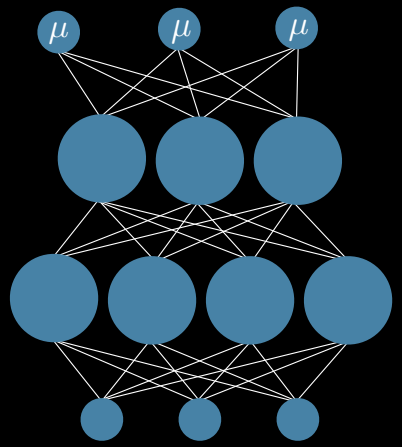
Concatenation cost $C(u_n, u_{n+1}) = \sum_{j=1}^{J} \left(u_{n+1}^{b}[j] - u_n^{e}]j]\right)^2$

Target cost $T(u_n) = \sum_{j=1}^{J} \left(u_n[j] - \mu_t[j]\right)^2$

~~Output of a parametric HMM synthesizer~~
current phone
preceding and succeeding phone

Linguistic Features

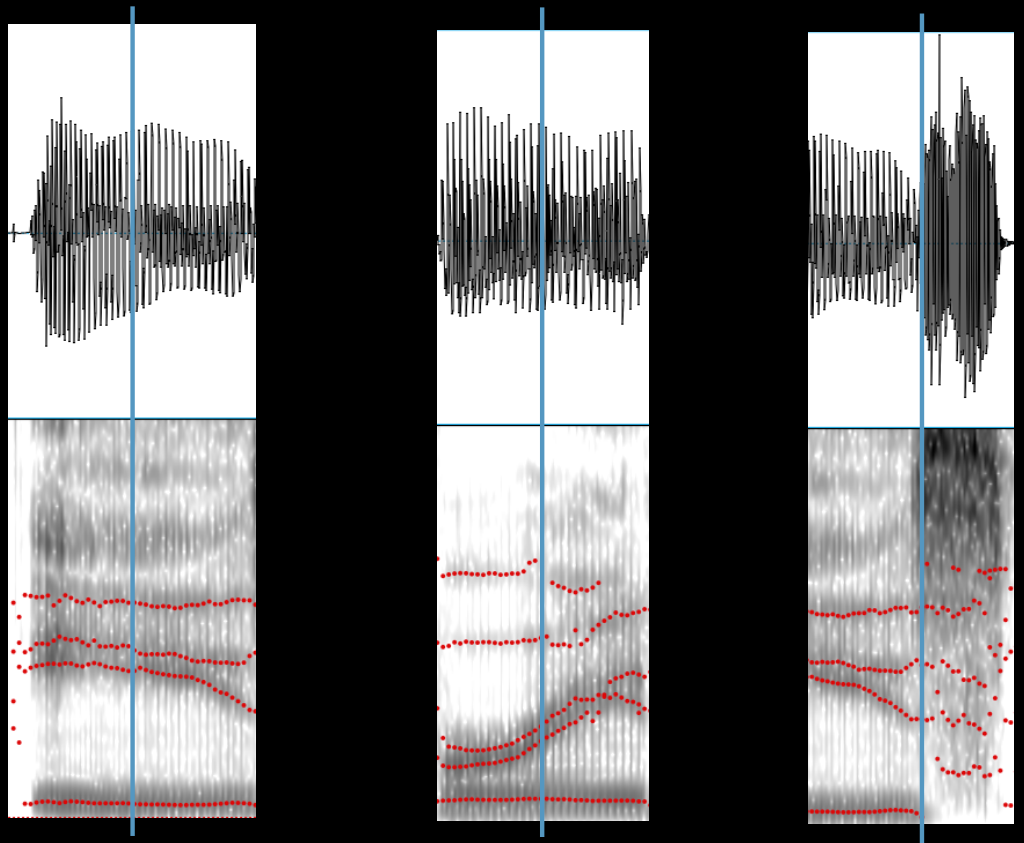
phoneme	•current phoneme •preceding and succeeding two phonemes •position of current phoneme within current syllable
syllable	•numbers of phonemes within preceding, current, and succeeding syllables •stress and accent of preceding, current, and succeeding syllables •positions of current syllable within current word and phrase •numbers of preceding and succeeding stressed syllables within current phrase •numbers of preceding and succeeding accented syllables within current phrase •number of syllables from previous stressed syllable •number of syllables to next stressed syllable •number of syllables from previous accented syllable •number of syllables to next accented syllable •vowel identity within current syllable
word	•guess at part of speech of preceding, current, and succeeding words •numbers of syllables within preceding, current, and succeeding words •position of current word within current phrase •numbers of preceding and succeeding content words within current phrase •number of words from previous content word •number of words to next content word
phrase	•numbers of syllables within preceding, current, and succeeding phrases •position of current phrase in major phrases
utterance	•numbers of syllables, words, and phrases in utterance



Input linguistic features

Deep MDN:
- 3 layers
- 512 hidden units
- ReLU

Role of deltas/variances at the concatenation



Stable formants:
zero Δ, small σ

Evolving formants:
non-zero Δ, small σ

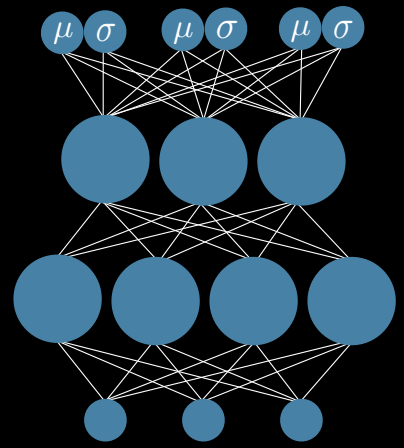
Transitions:
large Δ, large σ

Target cost

$$T(u_n, t) = \sum_{j=1}^{J} \frac{(u_n[j] - \mu_t[j])^2}{\sigma_t^2}$$

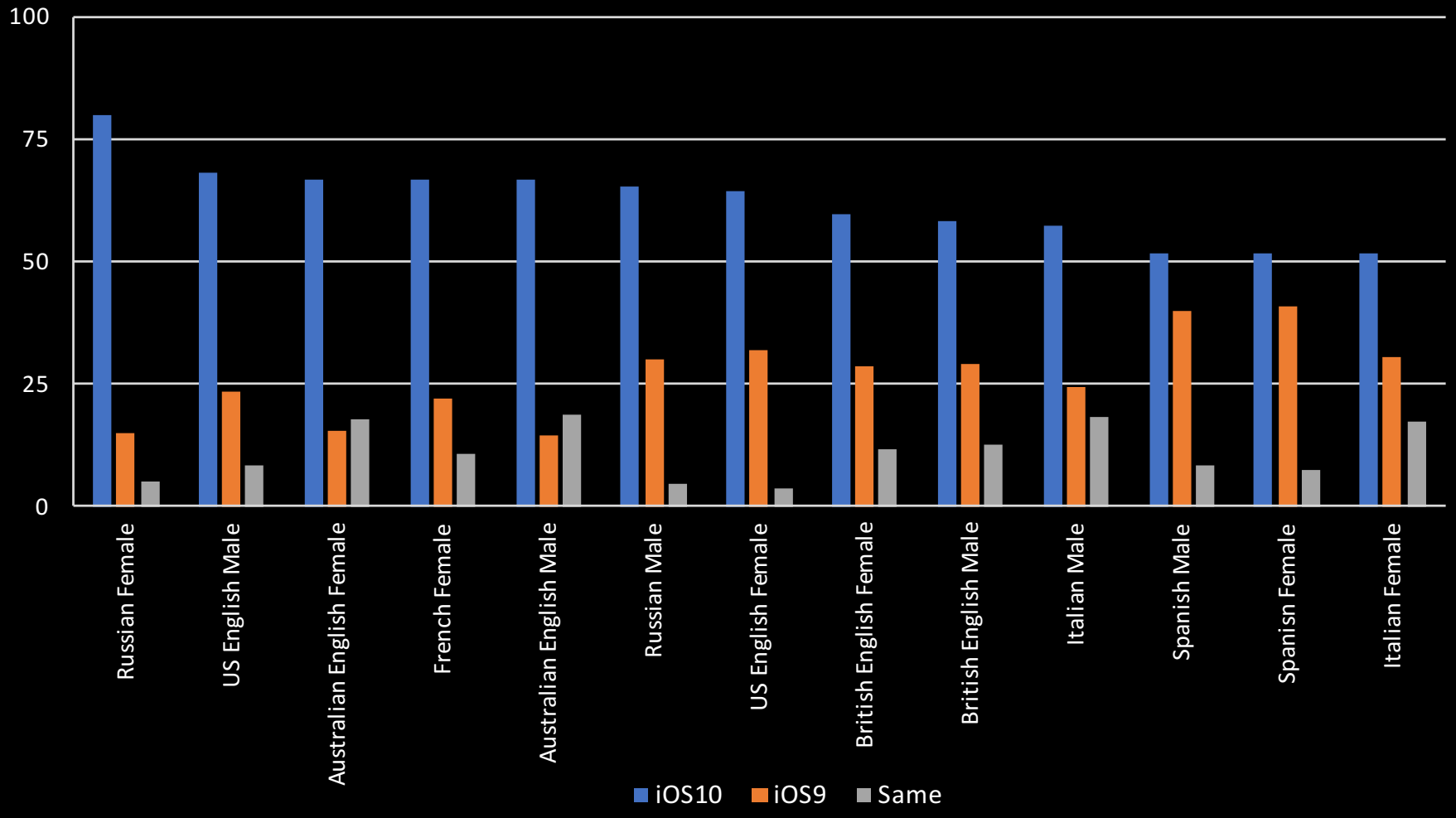
Concatenation Cost

$$C(u_n, u_{n+1}) = \sum_{j=1}^{J} \frac{(u_{n+1}^b[j] - u_n^e]j] - \delta_t[j])^2}{v_t^2}$$

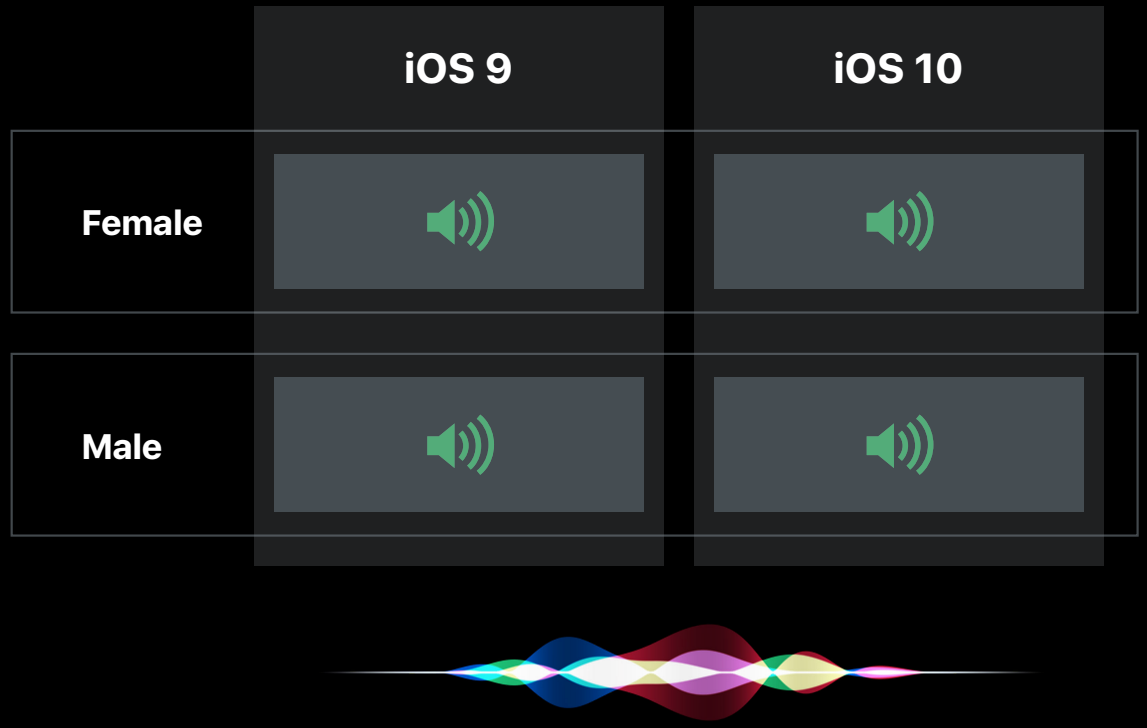


[1] C. Bishop, "Mixture density networks," Tech. Rep. NCRG/94/004, Neural Computing Research Group, Aston University, 1994.
[2] H. Zen, A. Senior, "Deep Mixture Density Networks for Acoustic Modeling in Statistical Parametric Speech Synthesis," ICASSP, 2014.

A/B Test (iOS10, iOS9, same)



iOS9 vs iOS10



iOS11 US female voice

