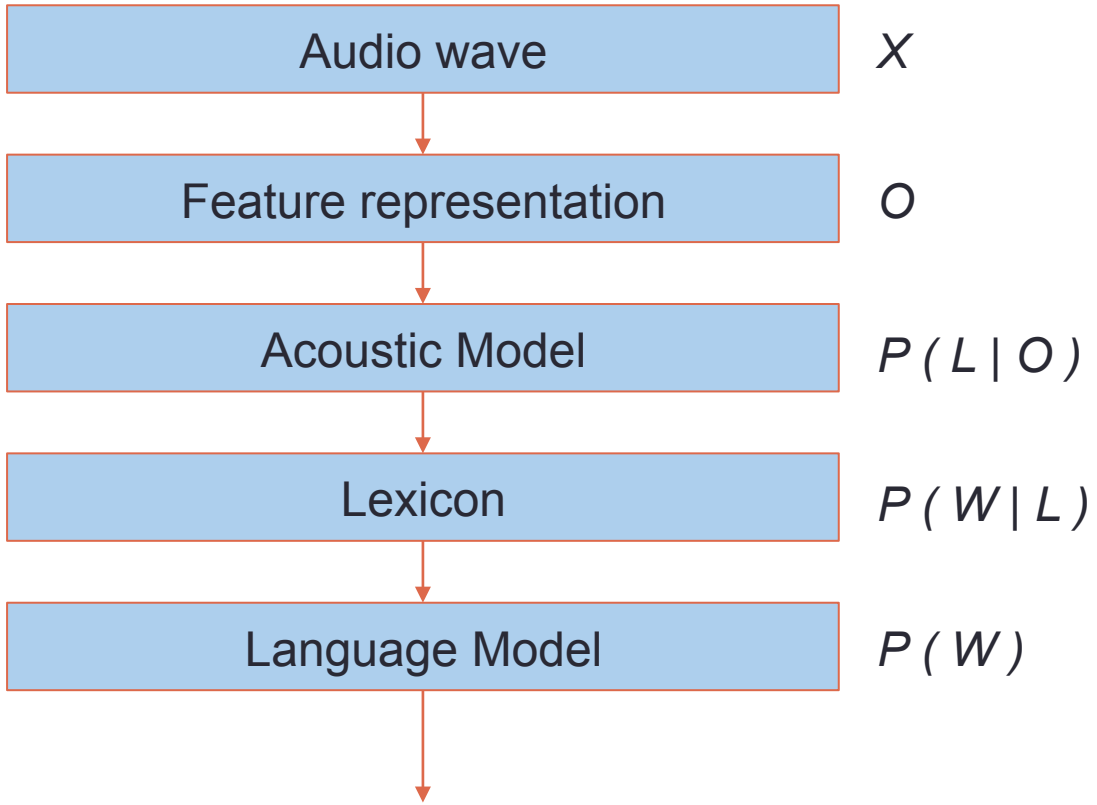
VARIOUS TOPICS IN SOUND PROCESSING

Recap

- DNN CNN RNN LSTM
- DNN in speech
- AM
  - Hybrid DNN-HMM
  - Bottleneck feature
- LM rescoring
- End-to-end
  - CTC

Traditional ASR pipeline



$$W^* = \text{argmax}_W P(W \mid X)$$

$$= \text{argmax}_{W,L} P(O \mid L) P(L \mid W) P(W)$$

Traditional ASR pipeline

- Traditional pipeline is easy to modify
  - Just swap LM for a new domain
  - New words can be added to lexicon
- Convoluted framework
  - Hard to understand the entire system
  - Each part has its own problems
- Each part is optimized individually
  - Each part built on top of (wrong) assumptions from previous step

Deep learning in ASR

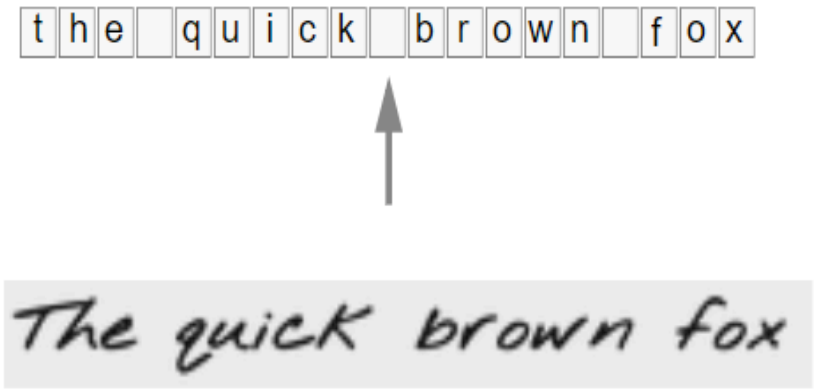
- Replace acoustic model $P(L \mid O)$ with deep learning
  - Hybrid DNN-HMM approach (previous lecture)
  - Improve acoustic model performance but still rely on other parts
- Replace language model with deep learning
  - Used for rescoring
  - Still the same limitations

End-to-end ASR

- Remove the pipeline
- Easier to maintain in terms of the pipeline
  - Less flexibility to mix-and-match
- Need more data
- More computation

End-to-end with CTC

- Connectionist Temporal Classification (CTC)



Handwriting recognition: The input can be (x, y) coordinates of a pen stroke or pixels in an image.

Speech recognition: The input can be a spectrogram or some other frequency based feature extractor.

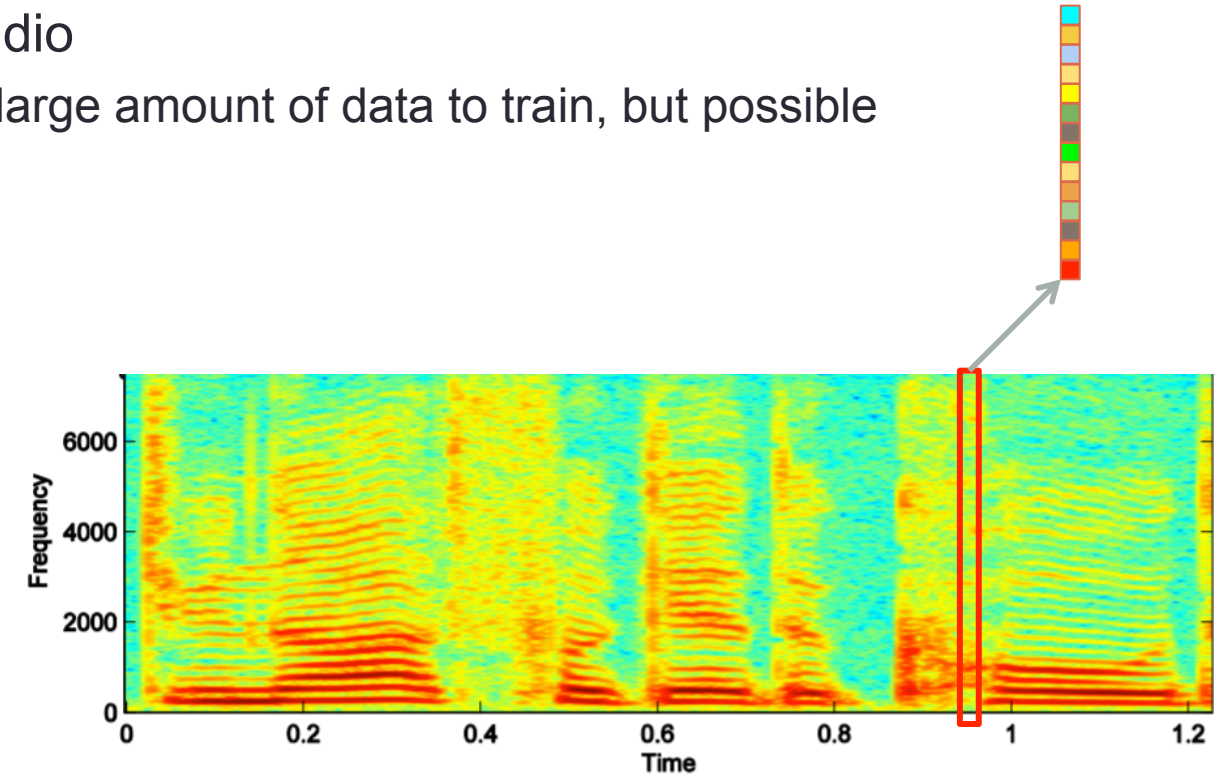
https://distill.pub/2017/ctc/

CTC Outline

- Preprocessing (features)
- CTC
  - Training
  - Decoding and LM

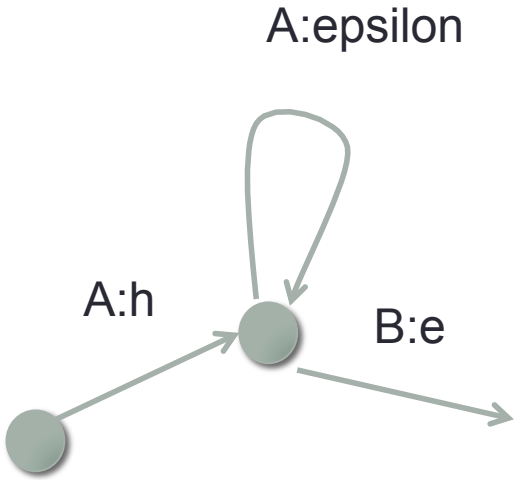
Preprocessing

- Two choices:
  - Spectrogram (filterbank) as inputs
    - Subsample (does not use every frame)
  - Use raw audio
    - Use very large amount of data to train, but possible

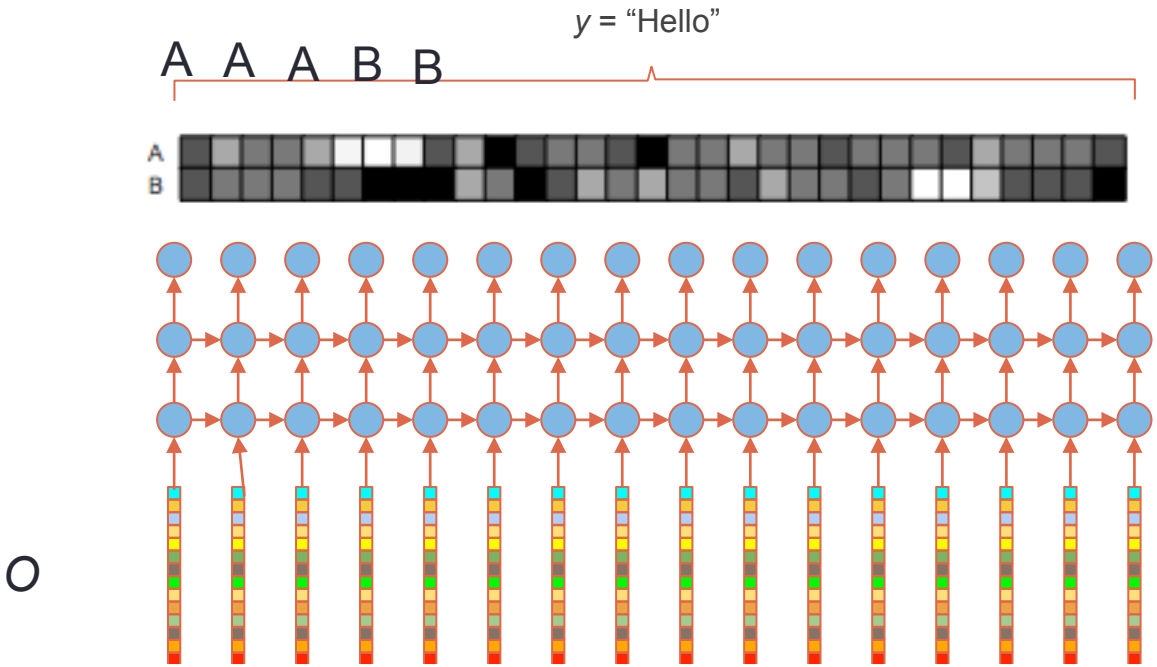


The problem with scale

- Letters in a sentence = y
- Frames in a sentence = x
- Length(x) >> length(x)



Hybrid DNN-HMM deals with it by using self transitions to eat up frames



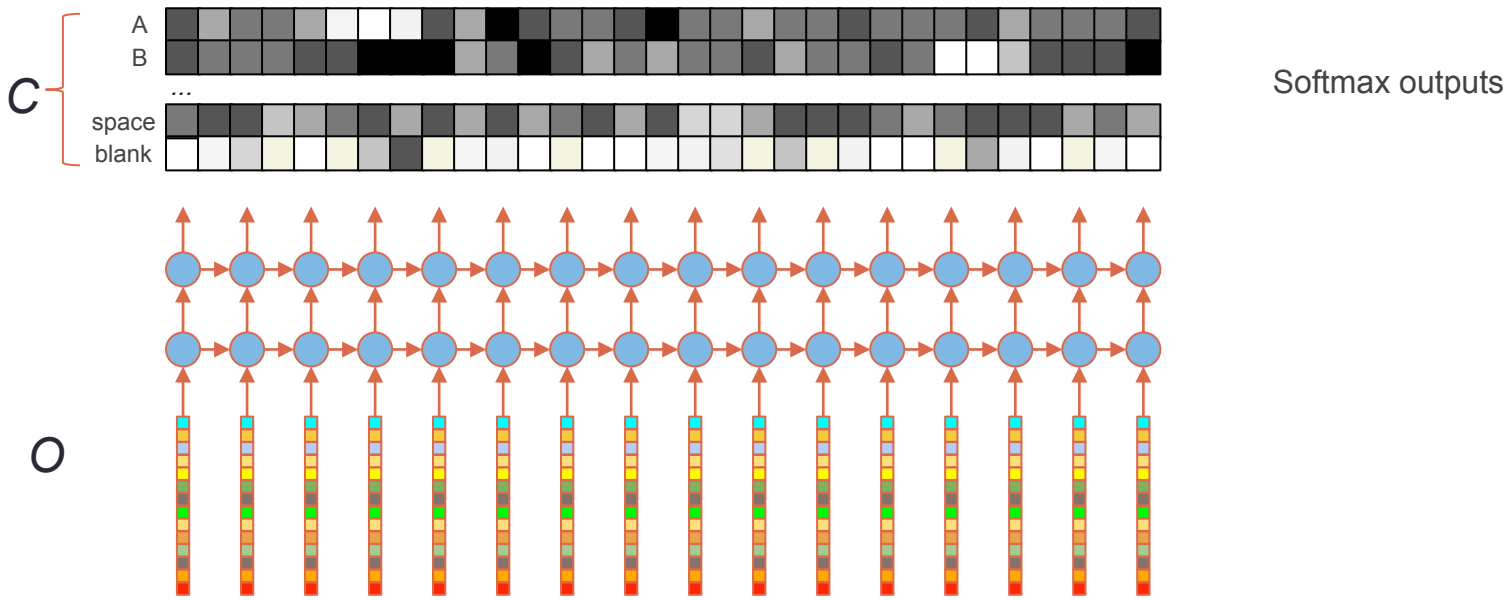
CTC

- Idea:
  - RNN output distribution over possible symbols ***c***, $c \in$ {A, B, C, …, Z, blank, space}
    - blank _ is an empty symbol
    - Space separate words
    - length(c) == length(x)
  - Define a mapping $\beta(c) \rightarrow y$
    - Remove repeats
    - Remove blanks
      - blank determine a new character C_C -> CC vs CC -> C

HHH_E_LLL_LO__	**=**	“HELLO”
HH_E_LLL_LO___	**=**	“HELLO”
YY_E_LLL_L_O_W	**=**	“YELLOW”
…		

CTC

- 1. RNN output distributions over possible symbols



CTC

- 1. RNN output distributions over possible symbols
- Assume independence over frames we get

$$P(c|x) = \prod_{i=1}^{N} P(c_i|x)$$

$P(c$ = HH_E_LLL_LO___$|x)$ = $P(c_1$ = H $|\ x)\ P(c_2$ = H $|\ x)\ P(c_3$ = blank $|\ x)$ … $P(c_{15}$ = blank $|\ x)$

CTC

- 2. Define a mapping function
  - Many sequence that yields the same transcription

P(c = HHH_E_LLL_LO__| *x*) = 0.1	“HELLO”
P(c = HH_E_LLL_LO___| *x*) = 0.03	“HELLO”
P(c = HH_E_L_LO_____| *x*) = 0.02	“HELLO”
P(c = YY_E_LLL_L_O_W| *x*) = 0.005	“YELLOW”
...	

All possibilities that yield the same output y

$$P(y \mid x) = \sum_{c:\beta(c)=y} P(c \mid x)$$

P(“HELLO”) = 0.1 + 0.03 + 0.02 + ...

CTC training

- Update network parameters ***θ*** to maximize likelihood of correct label ***y****

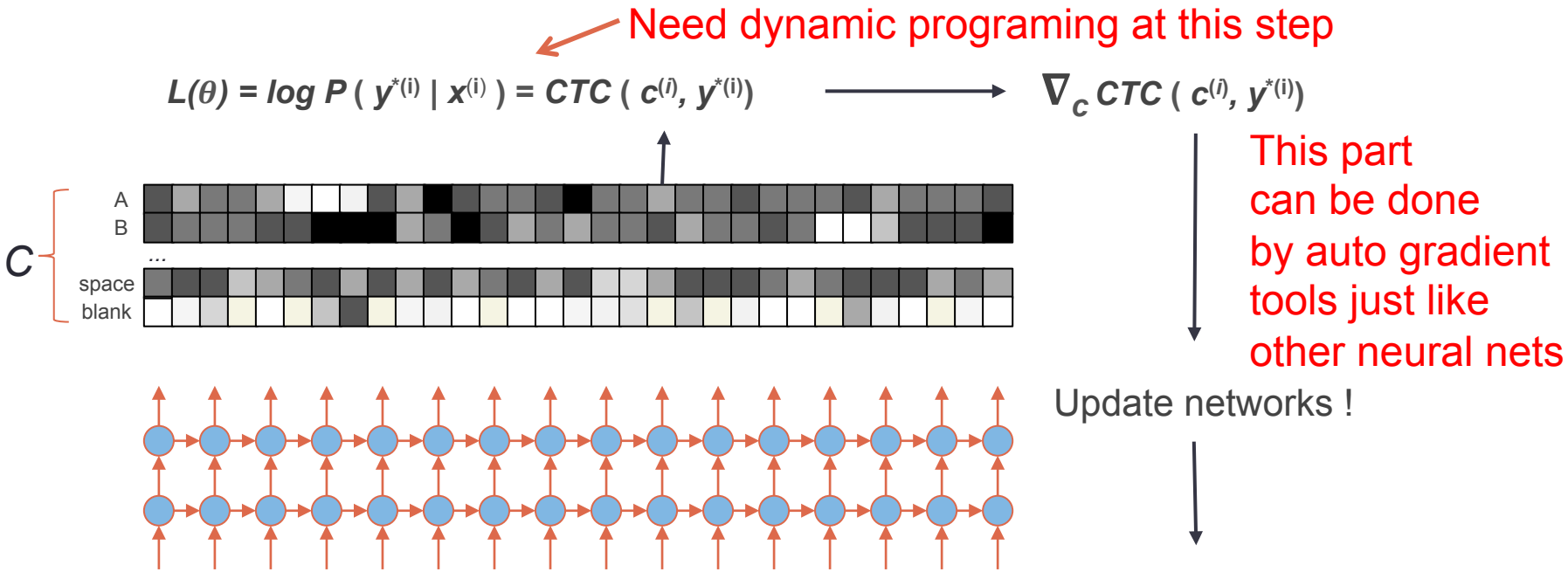
$$\theta^\star = \underset{\theta}{argmax} \sum_i log P(y^{\star(i)}|x^{(i)})$$

$$= \underset{\theta}{argmax} \sum_i log \sum_{c:\beta(c)=y^{\star(i)}} P(c|x^{(i)})$$

- [Graves et al., 2006] provides an efficient way to compute the inner summation and its gradient
  - Uses dynamic programming
- http://andrew.gibiansky.com/blog/machine-learning/speech-recognition-neural-networks/

CTC training

- Update network parameters θ to maximize likelihood of correct label y^*
- If we can compute the gradient, we can use gradient descent to optimize!



CTC decoding

- We have network output P(c|x) in every frame. Find the most likely transcription P(y|x)
  - Simple solution (bad): max decoding

$$\beta\left(\underset{c}{argmax} P(c|x)\right)$$

β("_____AAAAAA_____BBB_") = "AB"

  - Better solution: beamsearch (just like HMM decoding)
    - Keep track of many possible sequences
    - Prune lower scoring sequences

Downside of character-based methods

- Even with better decoding scheme, CTC model still makes spelling and linguistic mistakes

RNN output	Decoded Transcription
what is the weather like in bostin right now	what is the weather like in boston right now
prime miniter nerenr modi	prime minister narendra modi
arther n tickets for the game	are there any tickets for the game

Hannun et al., 2014

- Will require massive amounts of data (audio – transcription pair) to learn complicated spellings

"Try cough ski concerto"	"Tchaikovsky concerto"

- Traditional pipeline can use more data to train LM (just need text)

LM in CTC

- If we have word-base LM that give $P(w_{t+1} \mid w_1, w_2, w_3, \ldots w_t)$, we optimize:

$$\underset{w}{argmax} P(w|x) P(w)^{\alpha} [length(w)]^{\beta}$$

where ***P(w|x) = P(y|x)*** for every characters in ***w***

α and β are the same parameters as in traditional ASR pipeline

Beamsearch in CTC with LM

- Beamsearch to maximize:

$$\underset{w}{argmax} P(w|x)P(w)^{\alpha}[length(w)]^{\beta}$$

- Start with a set of candidates transcript prefix, A = {}

For t = 1..T:

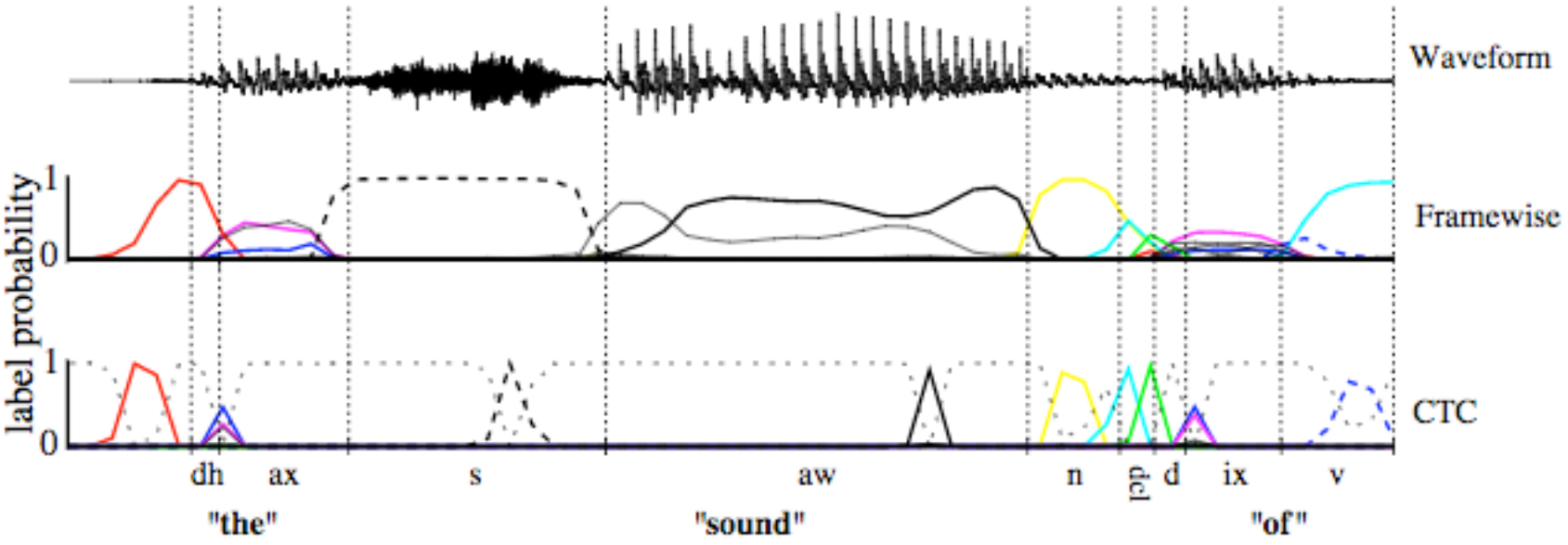
For each candidate in A, do:

New candidates are added to A_{new^*} by

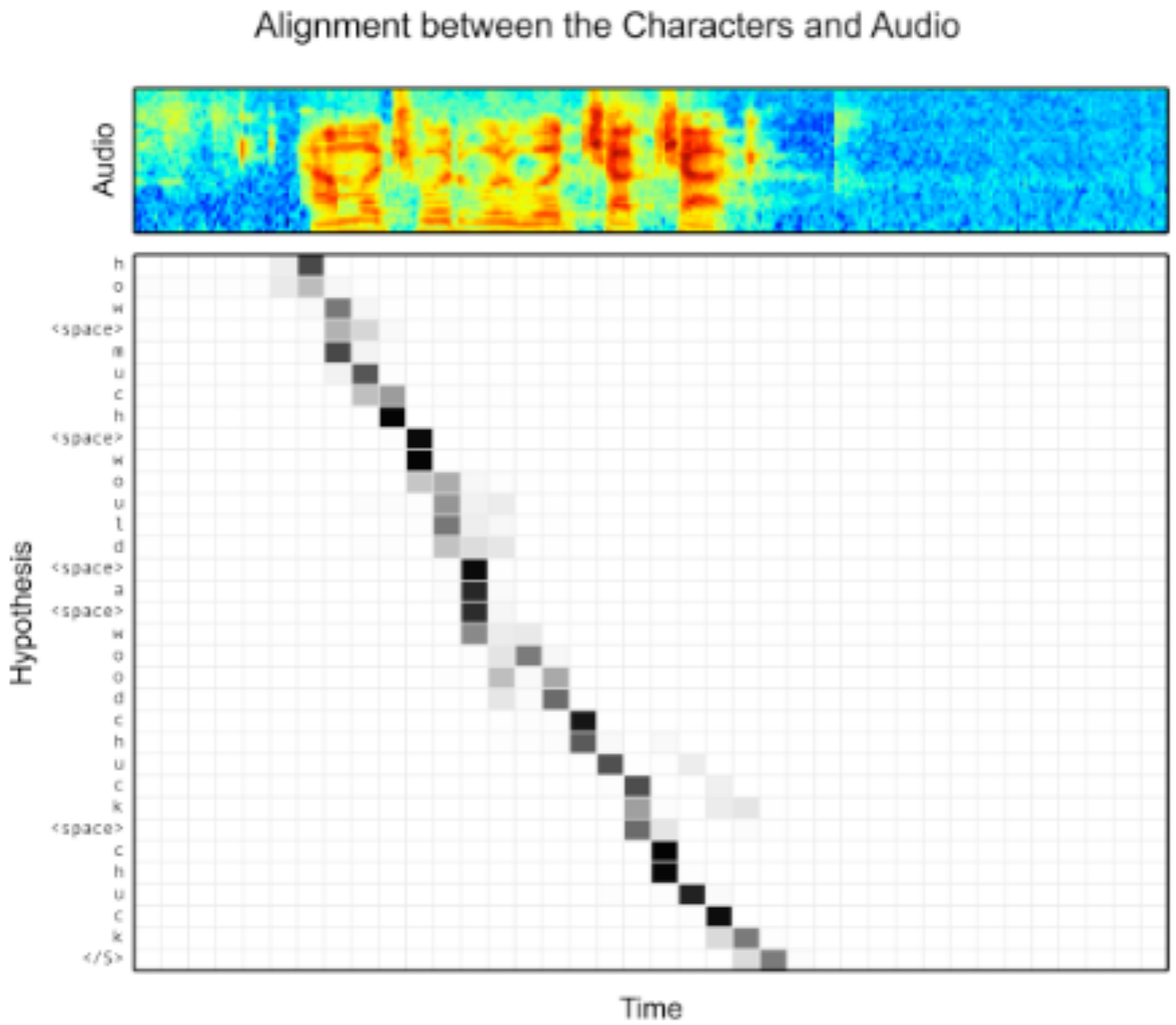
1. Add blank; prefix is still the same, update probability using AM
2. Add space to prefix; update probability using both AM and LM
3. Add a character to prefix; update probability using AM

A := ***K*** most probable prefixes in A_{new^*} (pruning)

CTC visualization



Graves, A. Connectionist Temporal Classification, 2006



Chan, W. Listen, Attend and Spell, 2015

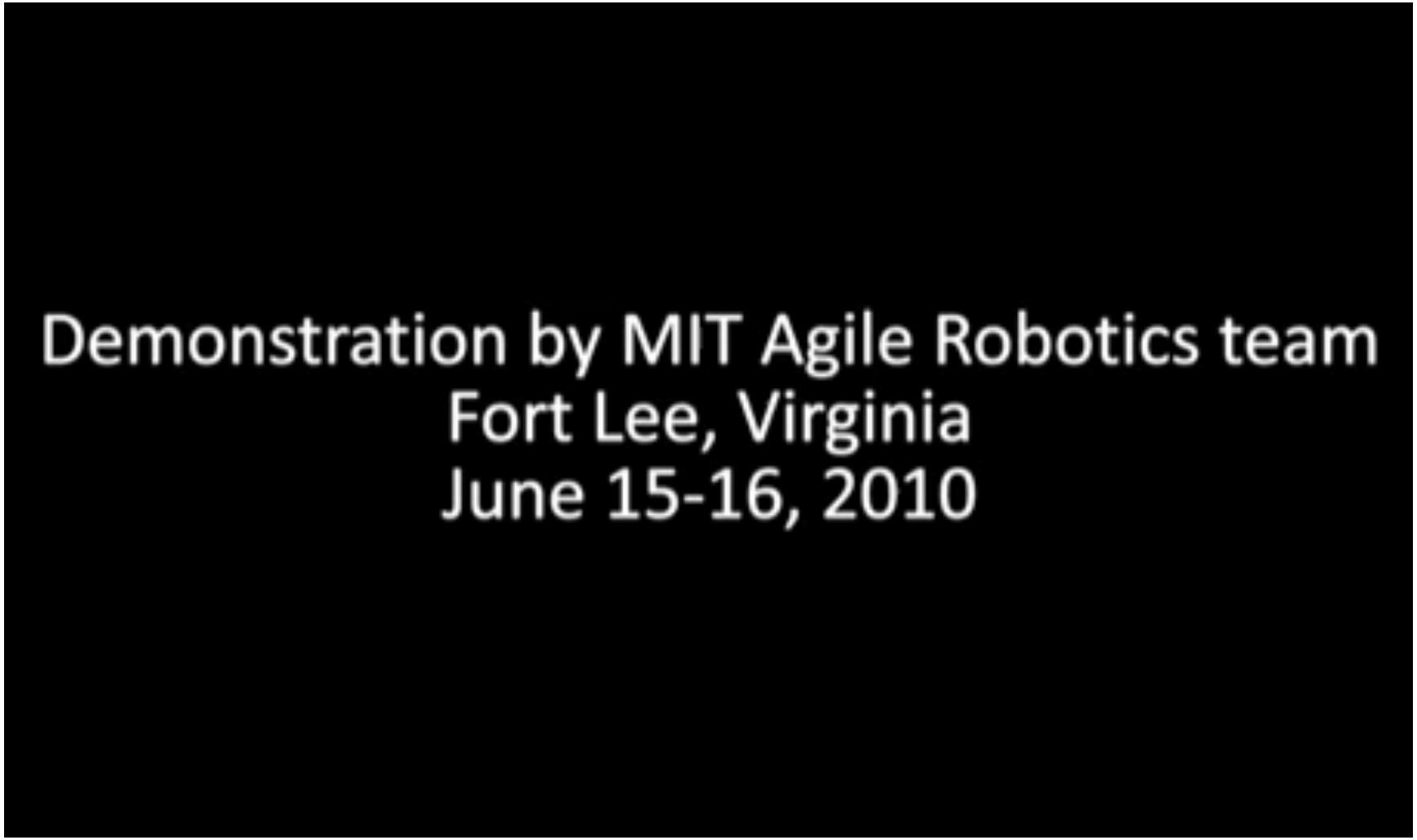
CTC in Thai (no LM)

- Original: ไม่ ได้ นึก ว่า เรื่อง ตลก นั้น มี หลาย ประเภท
- Iteration 1: กา ร ก ก ก ท่ ก น ป
- Iteration 43: ย นวา คงทรน จมีม กั
- Iteration 79: ไม่ ย นุน ว่า รง ตร นัน มีย ได่ ประบทน
- Iteration 115: ไม่ ได้ นึ คว่ารือง ตลนะง เมีไลา ประบพน
- Iteration 163: ไม่ ได้ย ไนึ่ว่า เเรือง ตัลนัน มี ไม ประเพ

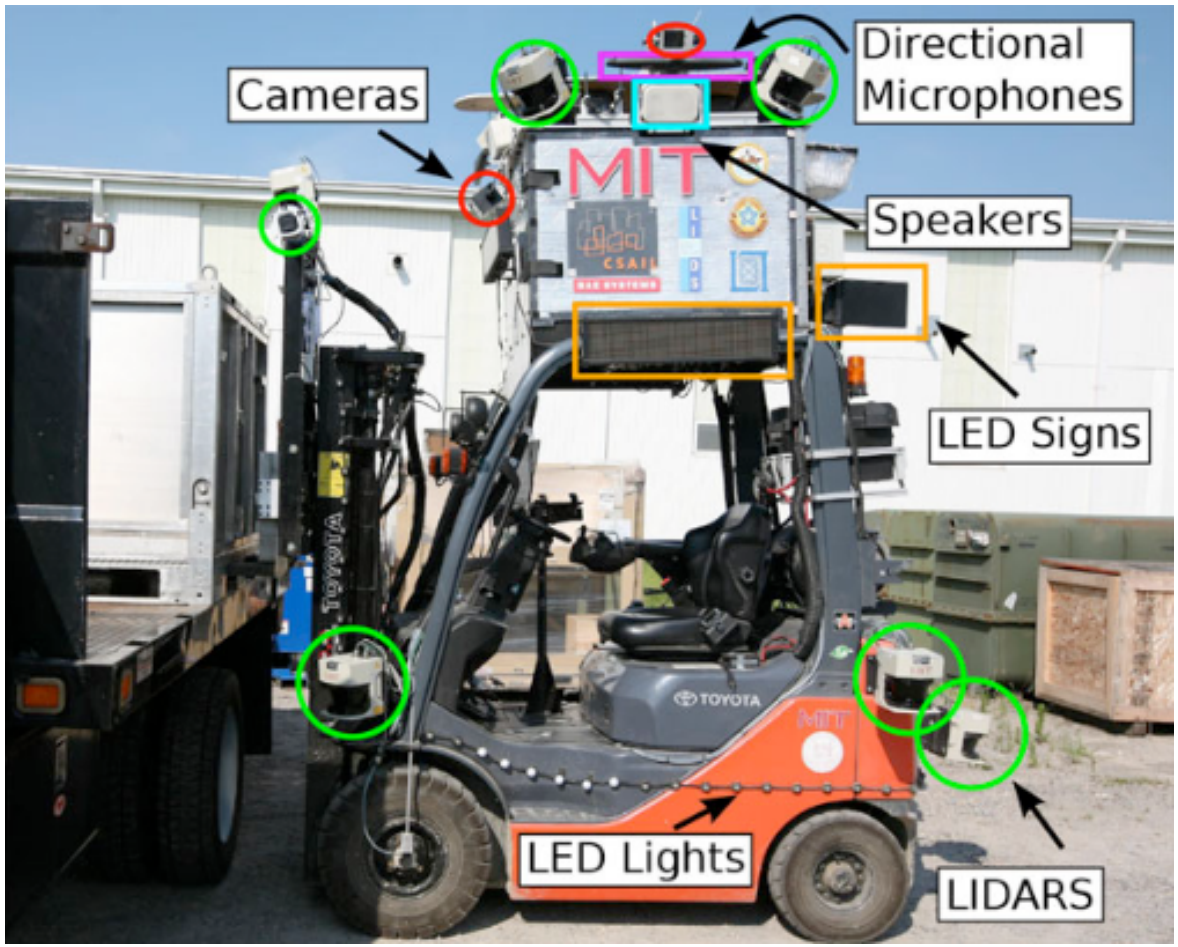
CTC in Thai (no LM)

- Original: ความ เปลี่ยนแปลง เหล่านี้ จะ สังเกต ได้ จาก ปริมาณ การ จัด แสดง ผลงาน หรือ นิทรรศการ ศิลปะ ของ ศิลปิน ทั้ง รุ่น เก่า และ รุ่น ใหม่ ใน กรุงเทพ
- Iteration 1: กา ก ก ก ก น ก ก ก่ ก ก่ ก่ ก ก่ ก่ ก่ ก
- Iteration 79: ความ เรีน ปร ลอ นี จั สน เป ได้ จา ปะริ มา การ กสลง พน นาน เวื สการ สึปะ ขอม สิ่ยปรีน ทาง รนตา นะรน ม ใน พน ที
- Iteration 163: ความ เป็ยน แปร รนี จั สงเป็ ได้ จอา ปัดิ มาร การ กสดง ผน มาน เหรือย ทสการ สินึปะ ของ สิ ลกปริน ทัง รุน ตั และ รน ใหม ใน ร เขก

Autonomous forklift voice command



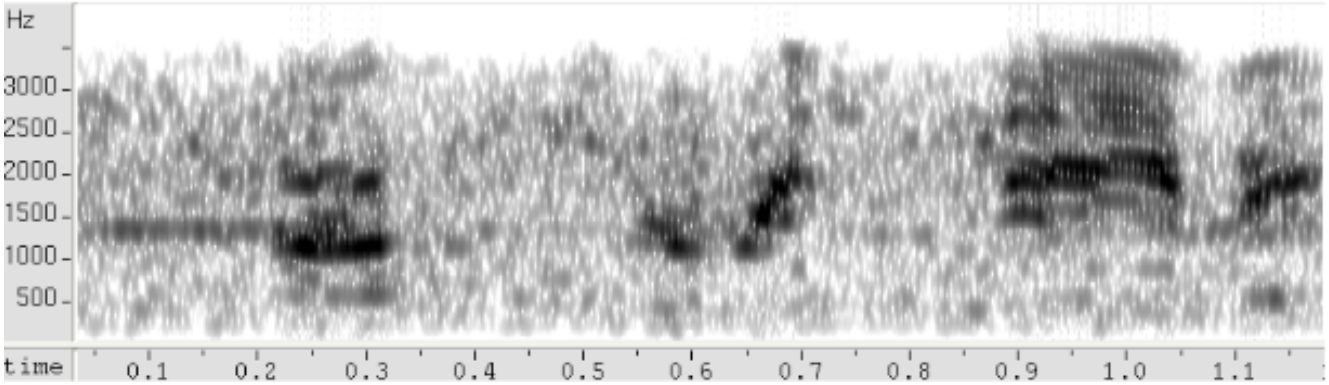
What techniques are needed?



M. Walter, "A Situationally Aware Voice-commandable Robotic Forklift Working Alongside People in Unstructured Outdoor Environments," 2015.

What techniques are needed?

- ASR
- Continuous listening
  - Is there speech?
  - Is it talking to me?
- Noise reduction
- Noise event detection, auditory scene analysis
- Identifying the speaker
- Speech interface are “natural,” should the commands be natural?
- Text2speech



Today topics

- VAD
- Noise reduction
  - Automatic Gain control
  - spectral subtraction
  - adaptive filtering
  - microphone array
  - Multi-condition training
- Language ID/Speaker ID/Emotion ID
  - i-vector
- Adaptation
  - VTLN
  - fMLLR
  - DNN adaptation
- Semi-supervised training and crowd sourcing
- Keyword search

Online vs offline processing

- Offline processing (or batch processing) – process the data as a whole
- Online processing – process the data as new data comes in
- Not to be confused with online and offline with regards to systems that interface with the internet to process data

Online processing version

- Mean removal

> 2. The information of the speech signal is in the fluctuation of the signal, and we can ignore the DC offset. The AFE standard applies the following filter to remove the DC offset from the waveform:
>
> $s_{of}[n] = s_{in}[n] - sin[n-1] + 0.999 s_{of}[n-1]$
>
> However, since we are processing the entire waveform at once instead of in real-time as it is being recorded, we can directly compute the DC offset of the entire utterance. What is the DC offset of this utterance? Remove the DC offset from the waveform.

- CMN
- Decoding
  - Output current best path which might change later
    - I. McGraw and A. Gruenstein, "Estimating Word-Stability During Incremental Speech Recognition," 2012

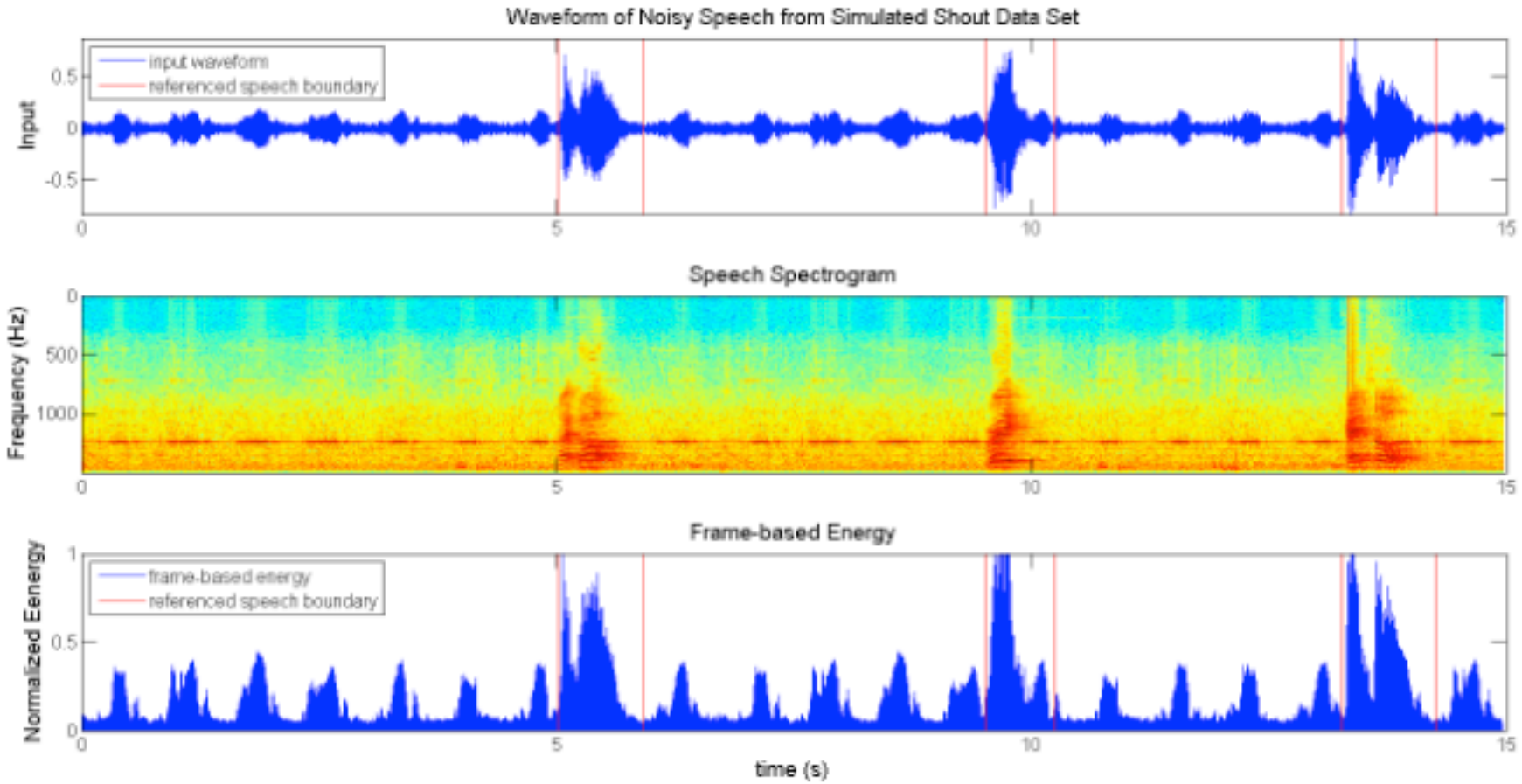
Real-time applications

- Real-time applications rely on online processing
- Real-time algorithms need to process data faster than the amount of input (RTF < 1)
  - Real-time factor (RTF), a 1 second speech takes 0.5 seconds to decode, RTF = 0.5
- Minimize latency (time from input until response from the user perspective)

Voice Activity Detection (VAD)

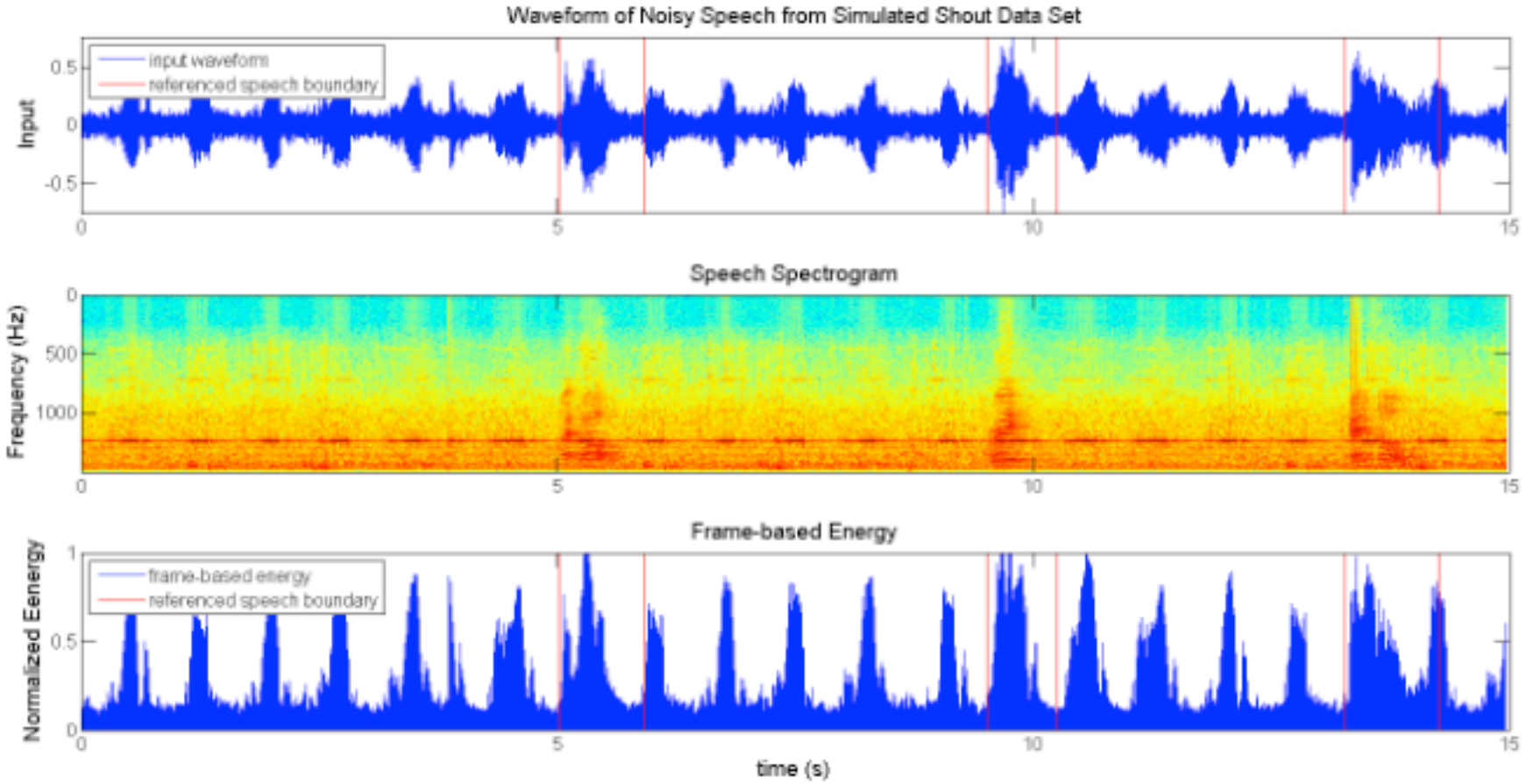
- or Speech Activity Detection (SAD)
- Determines if input frame is speech or non-speech
- Essential for every speech task!
- Feature-based vs ASR-based
- Features-based
  - Find good features that differentiate speech and non-speech
    - Energy, Harmonicity (pitch-detection), Spectral entropy, Modulation frequency, etc.
    - Best VADs used a combination of features
  - Requires less training data than ASR-based
  - Simple and fast to compute

Energy-based VAD



K. Luu, *Real-Time Noise-Robust Speech Detection, 2010*

Energy-based VAD in low SNR



WER with different VAD on forklift data

Energy-based

VAD	Words Correct (%)	WER (%)	FA (times/min)
Hand-labeled	56.3	45.0	0
RSE	44.1	59.6	0.70
AMR-VAD2	36.2	69.6	1.20
AFE	34.3	66.9	0.94
G.729B	7.0	93.5	0.30
MF+Harm	55.5	48.3	0.61

VAD can have a big effect on WER!
Push-to-talk button circumvent this, but not ideal.

E. Chuangsuwanich "Robust Voice Activity Detector for Real World Applications Using Harmonicity and Modulation Frequency," 2011.

ASR-based VAD

- Train ASR system to recognize speech and non-speech
- Two class speech, non-speech
- Can use DNN
- Powerful but slower than feature-based VAD
- Training script in available in kaldi
- Both methods usually have a post-processing step
  - Remove short speech/non-speech segments

Noise reduction

- Automatic gain control
- Spectral subtraction
- Adaptive filtering
- Microphone array
- Multi-condition training

Types noise

- 2 types of noise
  - Stationary noise (background noise)
    - Noise that is generated from a process with properties that are constant in time
  - Non-stationary noise (spontaneous noise)
    - Noise that is generated from a process with properties that are not constant in time
- Scale of observation is important
- Which noise is stationary?
  - White noise
  - 60 Hz electric hum
  - Beeping noise
  - Random hammer sounds
- Dealing with non-stationary noise is hard

Quantifying noise

- Signal-to-noise ratio (SNR) : compares the power of the signal to the noise

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

  - Measures in dB
  - 0 dB means noise and signal has equal power
  - >20 dB is considered clean speech
- To estimate SNR, usually requires VAD to identify speech/ non-speech region

Effect of noise on ASR

SNR/dB	Subway	Babble	Car	Exhibition	Average
clean	98.93	99.00	98.96	99.20	99.02
20	97.05	90.15	97.41	96.39	95.25
15	93.49	73.76	90.04	92.04	87.33
10	78.72	49.43	67.01	75.66	67.70
5	52.16	26.81	34.09	44.83	39.47
0	26.01	9.28	14.46	18.05	16.95
-5	11.18	1.57	9.39	9.60	7.93
Average between 0 and 20dB	69.48	49.88	60.60	65.39	**61.34**

Recognition accuracy on Aurora2 baseline

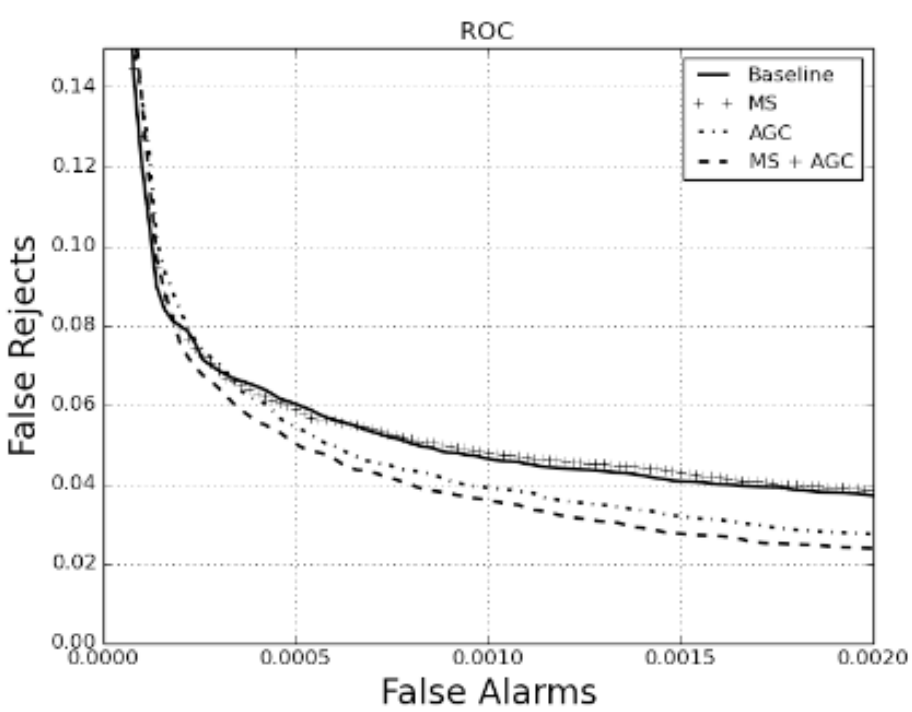
D. Pearce, The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions (2000)

Automatic Gain Control (AGC)

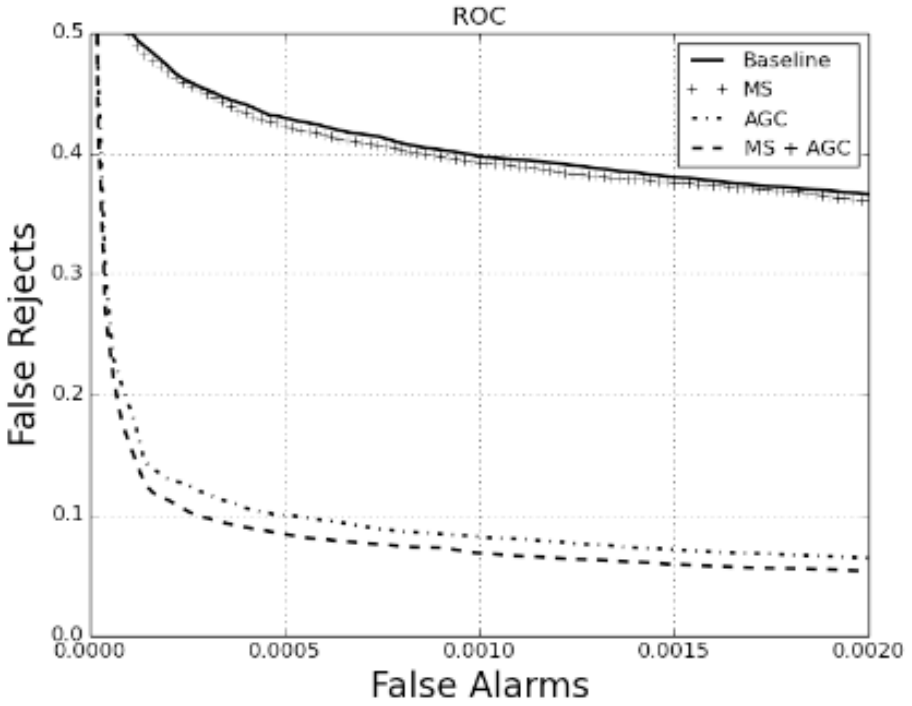
- MFCC features is not invariant of the gain (sound volume)
- Need to boost speech not noise
- Similar to how one would estimate SNR
  - Find speech regions
  - Estimate power
- Normalizing the gain can be easily done in an offline manner
- Tricky in online processing

R. Prabhavalkar, Automatic Gain Control and Multi-style Training for Robust Small-Footprint Keyword Spotting with Deep Neural Networks, 2015

Hotword with AGC



Clean test set

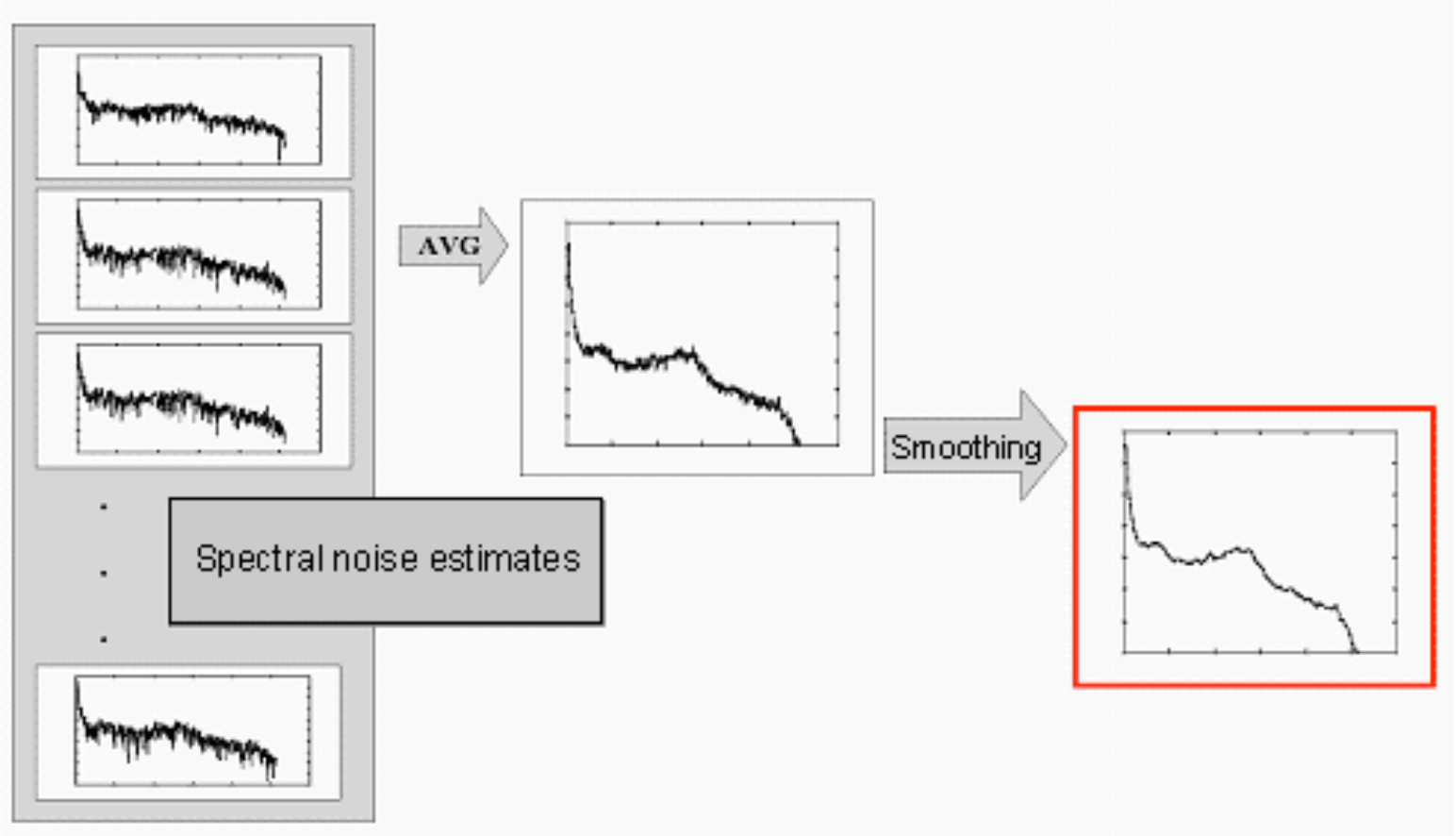


-5db car noise test set

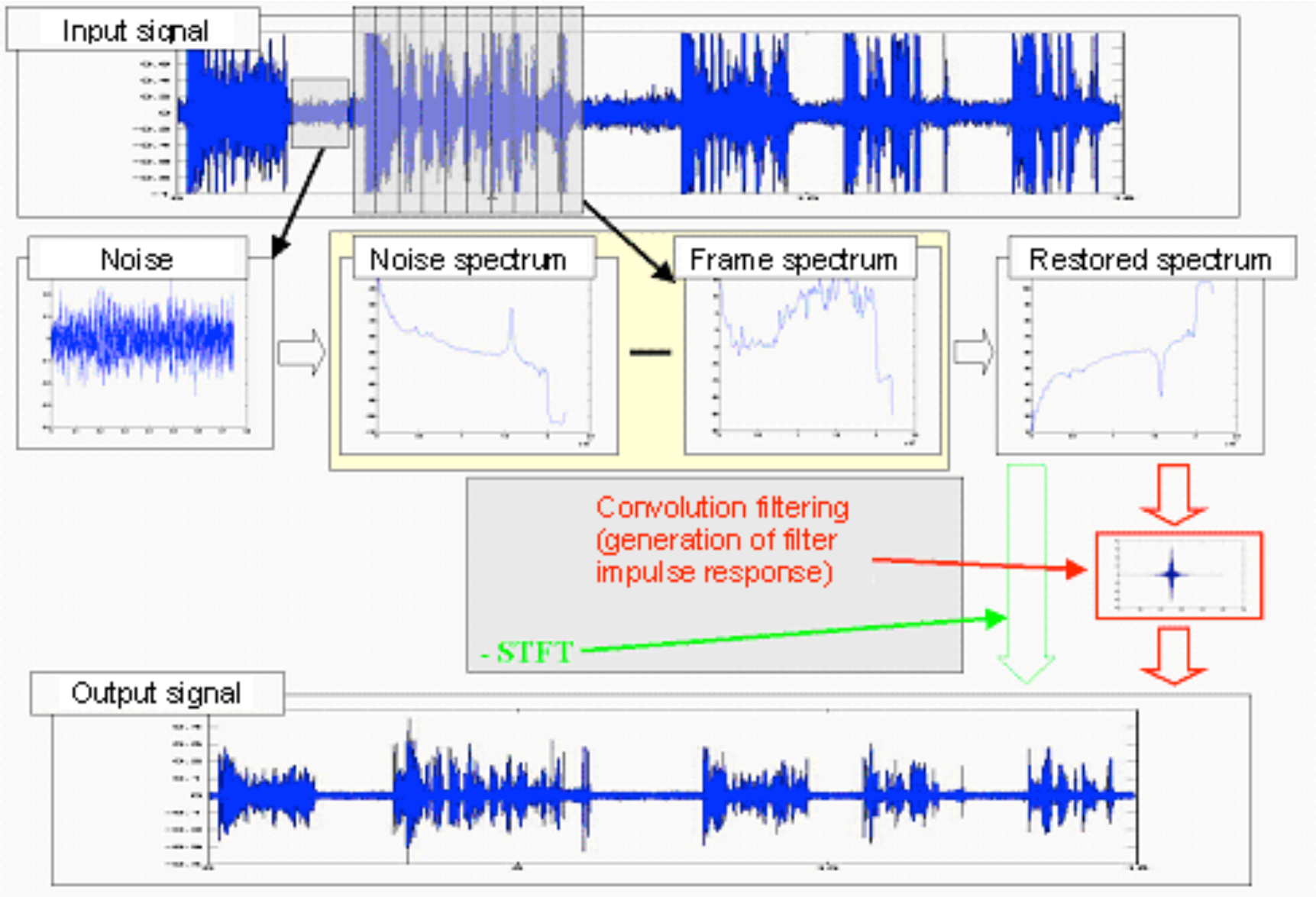
R. Prabhavalkar, Automatic Gain Control and Multi-style Training for Robust Small-Footprint Keyword Spotting with Deep Neural Networks, 2015

Spectral subtraction

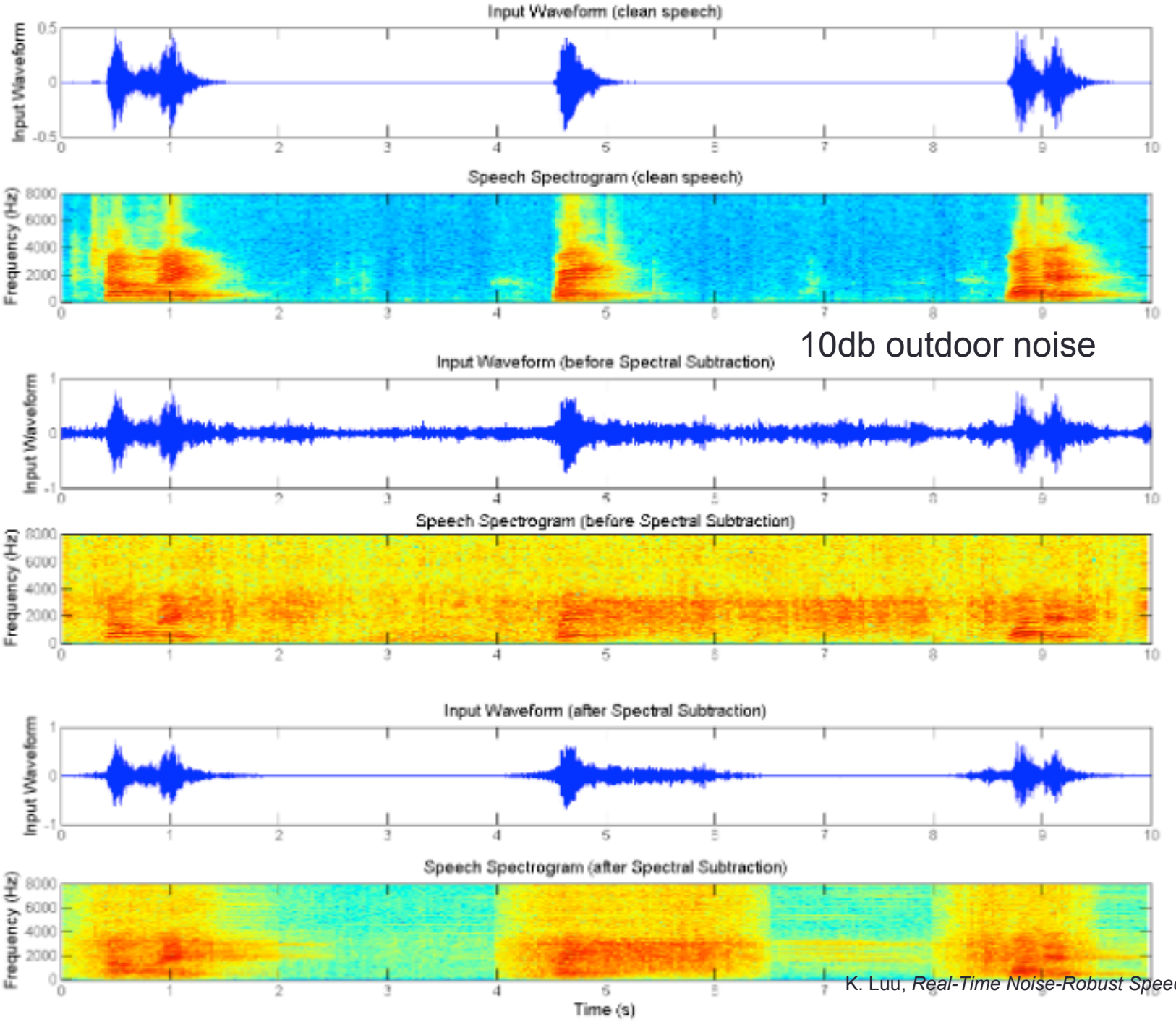
- Estimate average of noise spectrum and subtract from observed signal
- Good for stationary noise
- For some type of non-stationary noise, update noise estimate more frequently
- Assumes start of recording is noise only
  - Otherwise, need VAD



https://sound.eti.pg.gda.pl/denoise/noise.html



https://sound.eti.pg.gda.pl/denoise/noise.html



K. Luu, *Real-Time Noise-Robust Speech Detection, 2010*

Effect on ASR

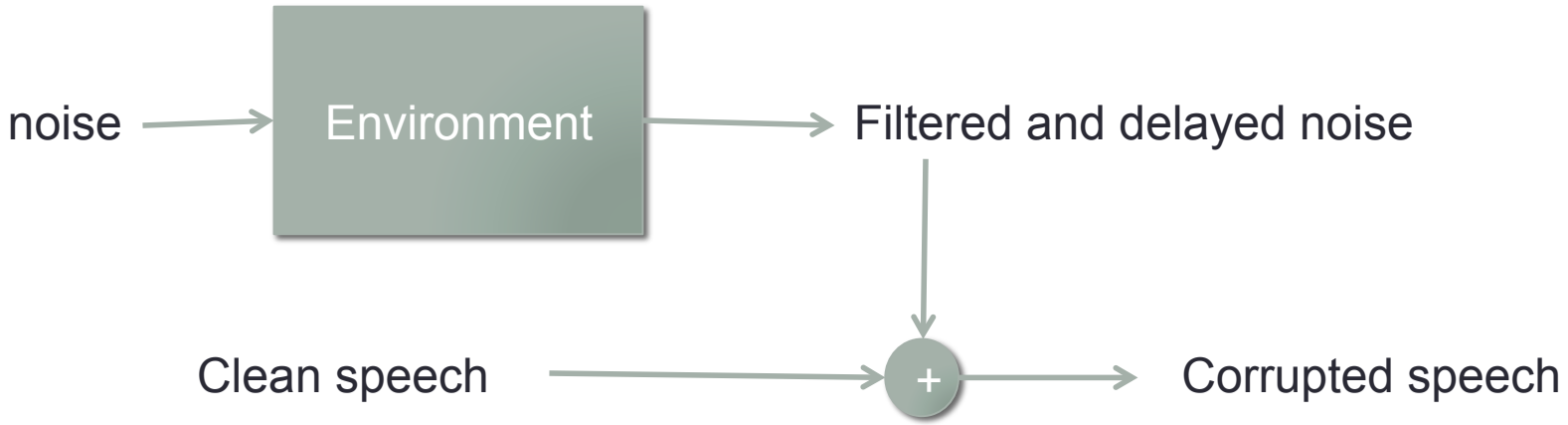
Front-end Configuration	Clean HMM Set	Multi-condition HMM Set
No compensation	65.01	86.21
SS only	75.64	89.07

Results on Aurora II

E. Choi, *Noise Robust Front-end for ASR using Spectral Subtraction, Spectral Flooring and Cumulative Distribution Mapping*. 2004

Adaptive filtering

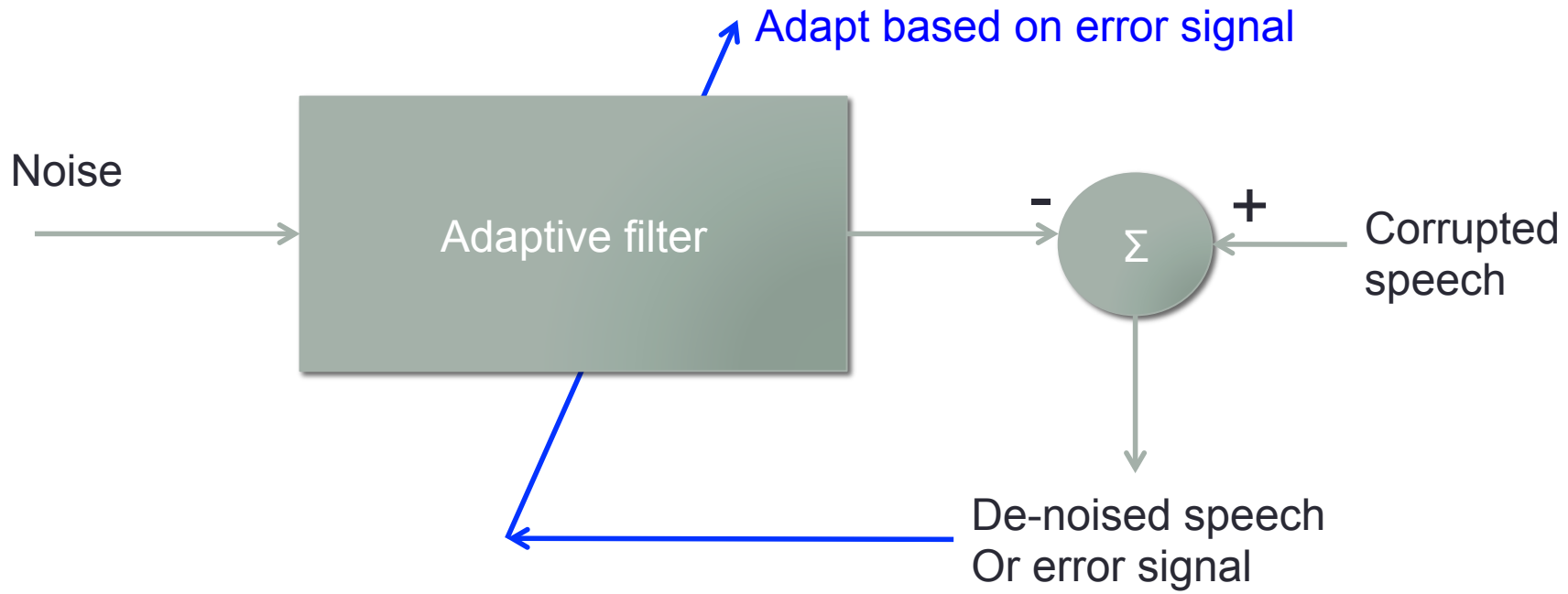
- If you know the interfering noise, you can subtract it from the signal.
  - Car speech + noise from radio
- But if the noise is corrupted, you cannot directly subtract



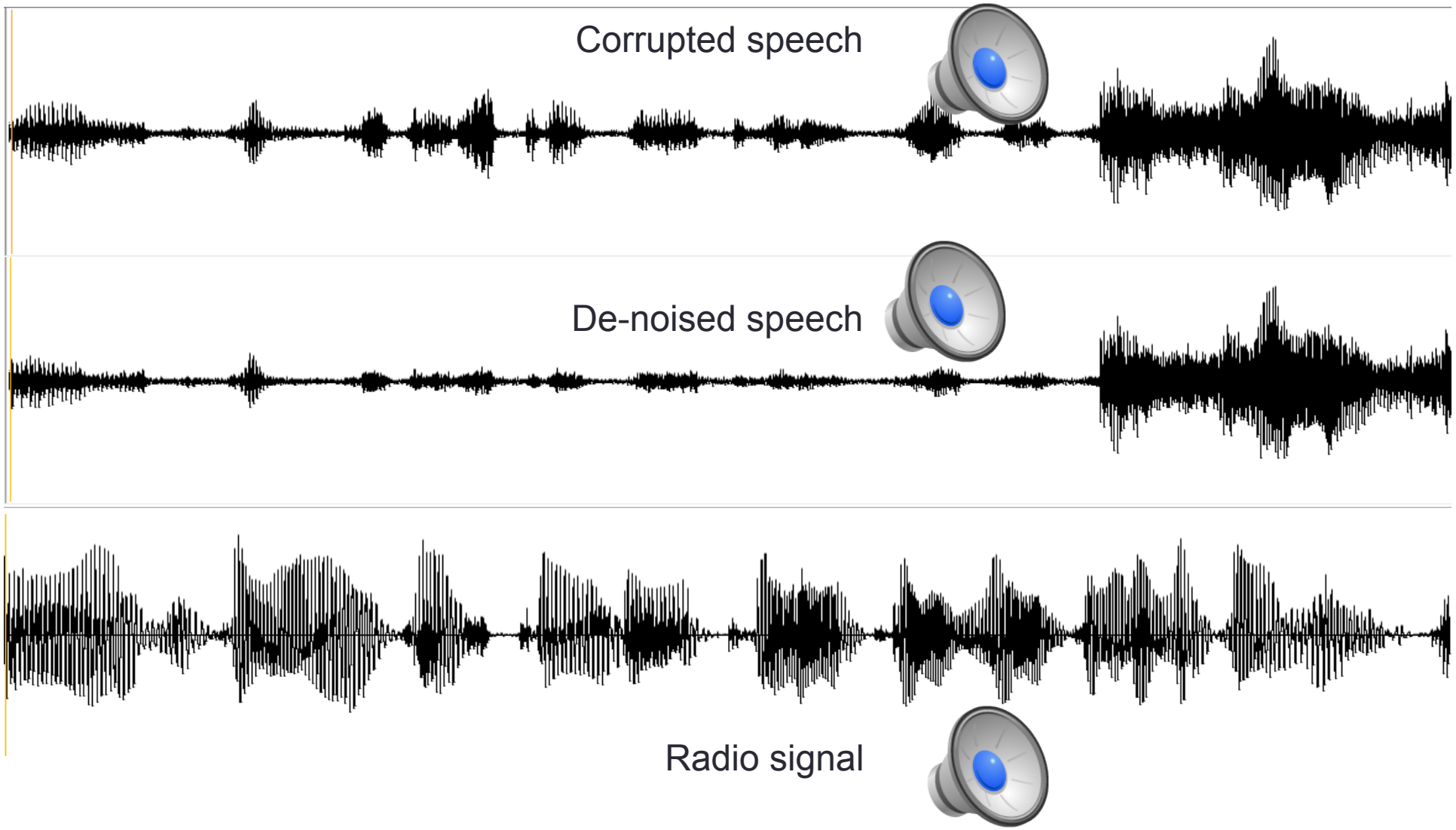
- Can we estimate the environment filter?

Adaptive filtering

Assumption: Noise and signal are uncorrelated

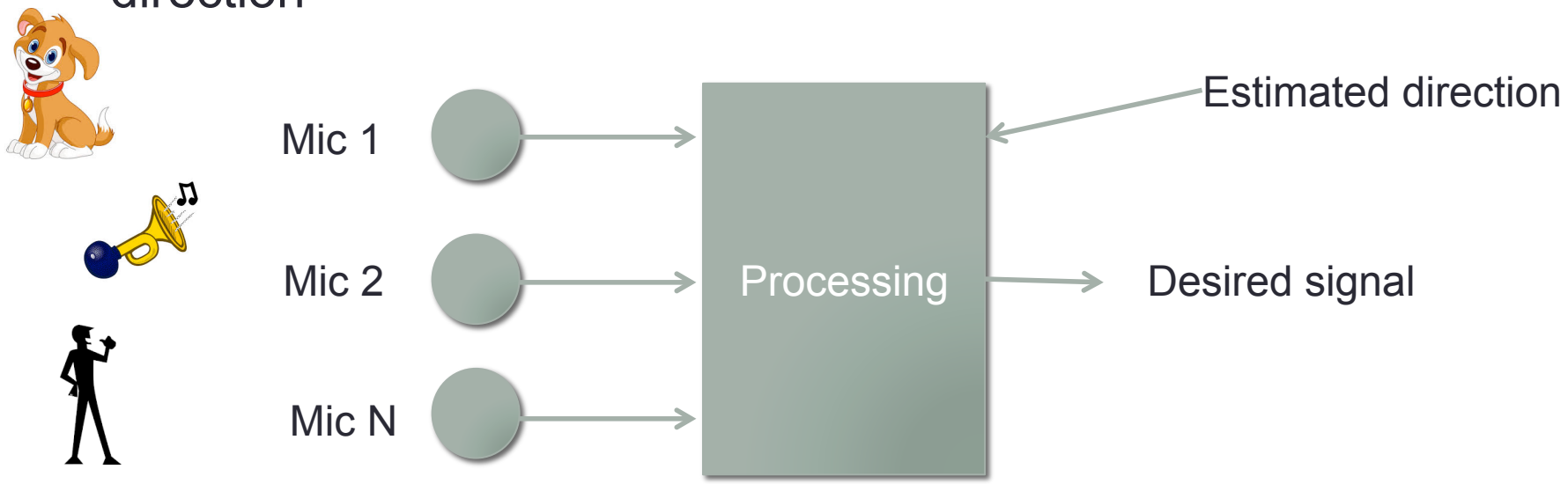


Goal: Minimize the power of the error signal
by changing the filter
Estimate and update chunk by chunk


Corrupted speech
De-noised speech
Radio signal

Microphone array/ beamforming

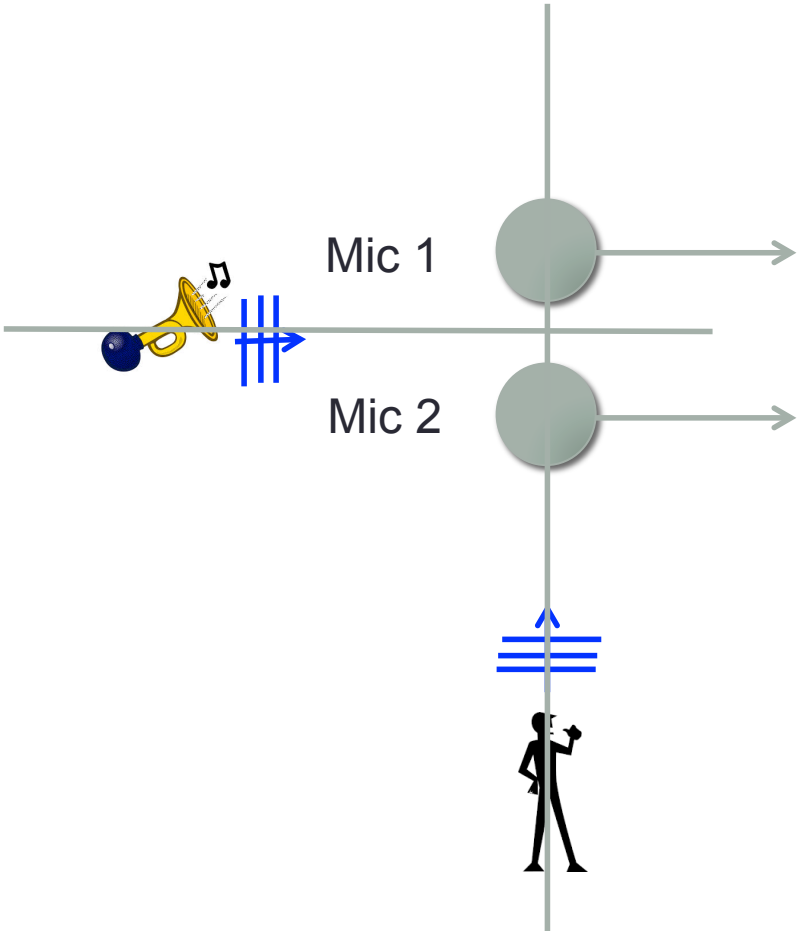
- Human has two ears which help localize the sound from different locations
- Microphone array steers the microphone to listen from a certain direction



- Theoretically, need N microphones to remove N-1 noise directions

Delay and sum beamforming

- Consider two microphones separated by 4 cm



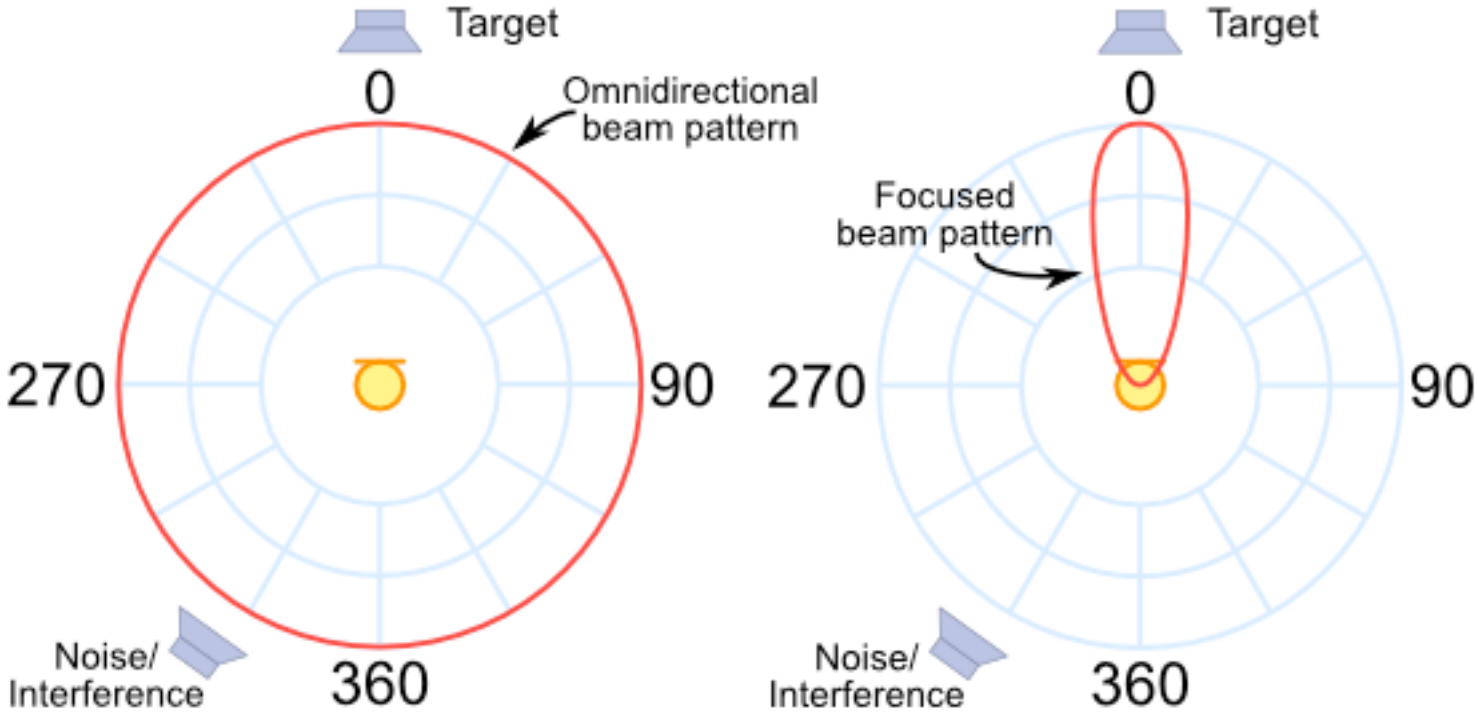
Mic 1 will listen to the speech
0.04/343 seconds later than mic 2

Which is 0.01/343 * 16000
= 1.88 samples later

Delay the signal by 1.88
and sum the signal from the two
microphones will boost the signal from
that direction

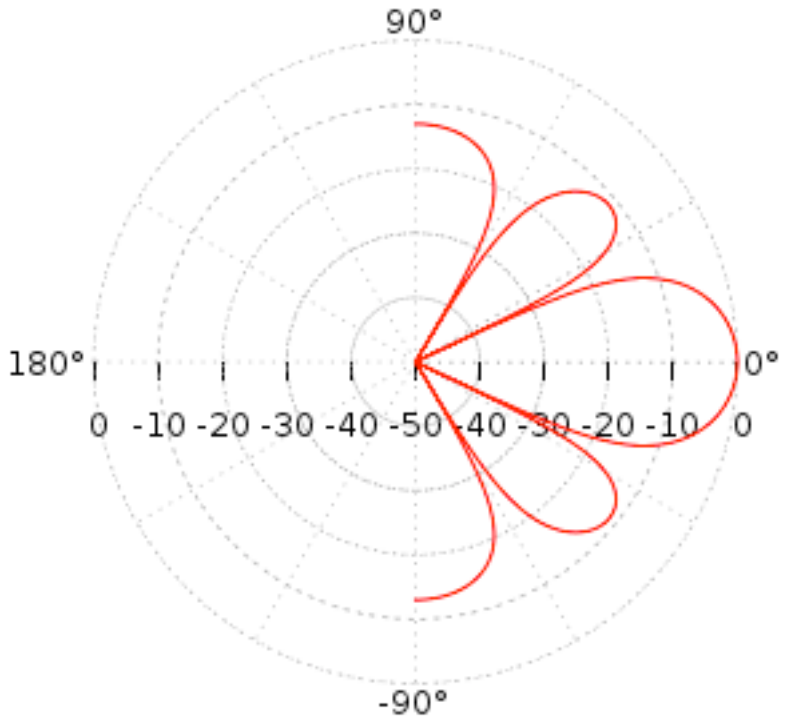
Beam pattern

- The capability of the array is usually summarized by a beam pattern diagram



http://www.labbookpages.co.uk/audio/beamforming/delaySum.html

Beam pattern



More microphones in an array lets you put more nulls in the beam pattern.

http://www.labbookpages.co.uk/audio/beamforming/delaySum.html

Multi-condition training

- Train using data with different kinds of corruption
  - Create more by data by adding different kind of noise at different SNR
- The model becomes more robust
  - But sometimes perform worse in clean situation
- Used in DNN training to create more data (data augmentation)

SNR/dB	Subway	Babble	Car	Exhibition	Average
clean	98.93	99.00	98.96	99.20	99.02
20	97.05	90.15	97.41	96.39	95.25
15	93.49	73.76	90.04	92.04	87.33
10	78.72	49.43	67.01	75.66	67.70
5	52.16	26.81	34.09	44.83	39.47
0	26.01	9.28	14.46	18.05	16.95
-5	11.18	1.57	9.39	9.60	7.93
Average between 0 and 20dB	69.48	49.88	60.60	65.39	**61.34**

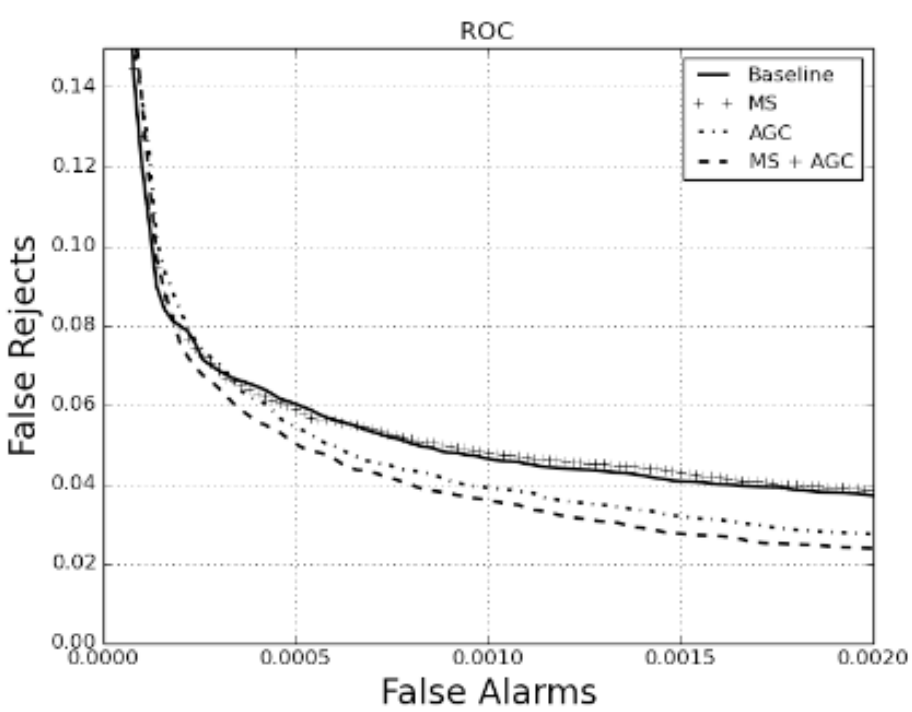
Training with clean data only

Multi condition

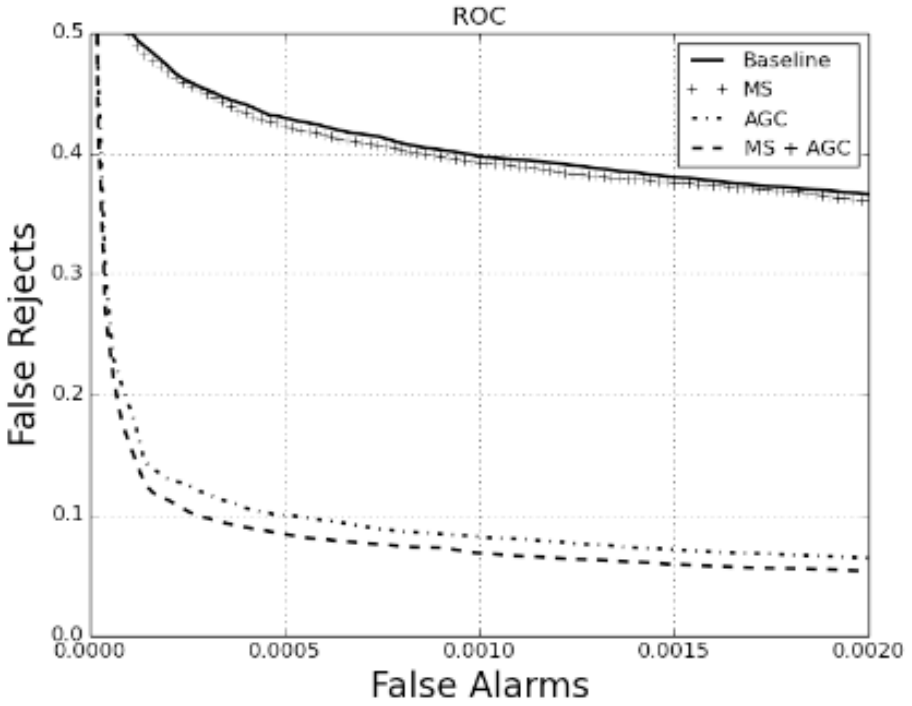
SNR/dB	Subway	Babble	Car	Exhibition	Average
clean	98.68	98.52	98.39	98.49	98.52
20	97.61	97.73	98.03	97.41	97.69
15	96.47	97.04	97.61	96.67	96.94
10	94.44	95.28	95.74	94.11	94.89
5	88.36	87.55	87.80	87.60	87.82
0	66.90	62.15	53.44	64.36	61.71
-5	26.13	27.18	20.58	24.34	24.55
Average between 0 and 20dB	88.75	87.95	86.52	88.03	**87.81**

D. Pearce, The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions (2000)

Hotword with AGC+MC



Clean test set



-5db car noise test set

R. Prabhavalkar, Automatic Gain Control and Multi-style Training for Robust Small-Footprint Keyword Spotting with Deep Neural Networks, 2015

Data augmentation

- There are other kinds of variation in speech
  - Speaking rate
  - Pitch
- Free gain in performance!

System	Fold	Epochs	LM	SWB	CHE	Total
Baseline	1	6	fg	13.7	27.7	20.7
VTLP	3	2	fg	13.1	26.5	19.9
VTLP	5	2	fg	13.2	26.7	20.0
VTLP + time-warp	3	2	fg	13.3	26.8	20.1
Tempo-perturbed	3	2	fg	13.5	27.0	20.3
Speed-perturbed	3	2	fg	13.1	26.1	19.7
Speed-perturbed	3	6	fg	12.9	25.7	19.3

T. Ko, Audio Augmentation for Speech Recognition, 2015

Today topics

- VAD
- Noise reduction
  - Automatic Gain control
  - spectral subtraction
  - adaptive filtering
  - microphone array
  - Multi-condition training
- Language ID/Speaker ID/Emotion ID
  - i-vector
- Adaptation
  - VTLN
  - fMLLR
  - DNN adaptation
- Semi-supervised training and crowd sourcing
- Keyword search

Speaker identification/verification

- Task: Given an utterance
  - Identification: who is this speaker?
  - Verification: is this spoken by X?

- Identification task often assume the unknown voice must come from a set of known speaker – a closed-set identification

Much content in the speaker ID section are adapted, with permission, from N. Dehak "Low-dimensional Speech Representation Based on Factor Analysis and its Applications," *Interspeech Tutorial*, 2011

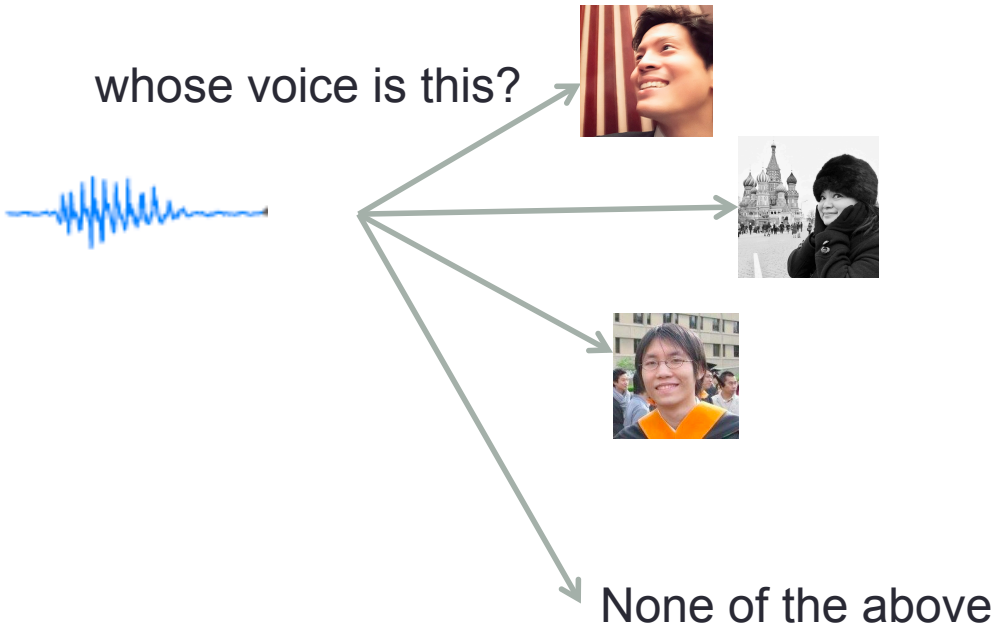
Speaker verification/authentication/ detection

- Determine whether unknown speaker matches a specific speaker

- Unknown speaker can come from a set of unknown speakers – an open-set verification

Open-set identification

- Adding “none of the above” option turns close-set identification into open-set identification



Speaker Diarization

- Who speaks when?
- Determine when a speaker change occurs (segmentation)
- Group together segments belonging to the same speaker (clustering)
- The speakers in the conversation can be unknown to the system

Speech modalities

- Two variants
  - Text dependent : must say the phrase
  - Text independent : can say anything
- which to use depends on the nature of the application

Text dependent

- Gowajee
- Application with strong control over user input
- Knowledge of spoken word helps improve performance

Text independent

- Application with less control over user input
- More flexible, but also harder problem
- ASR can provide info about spoken text

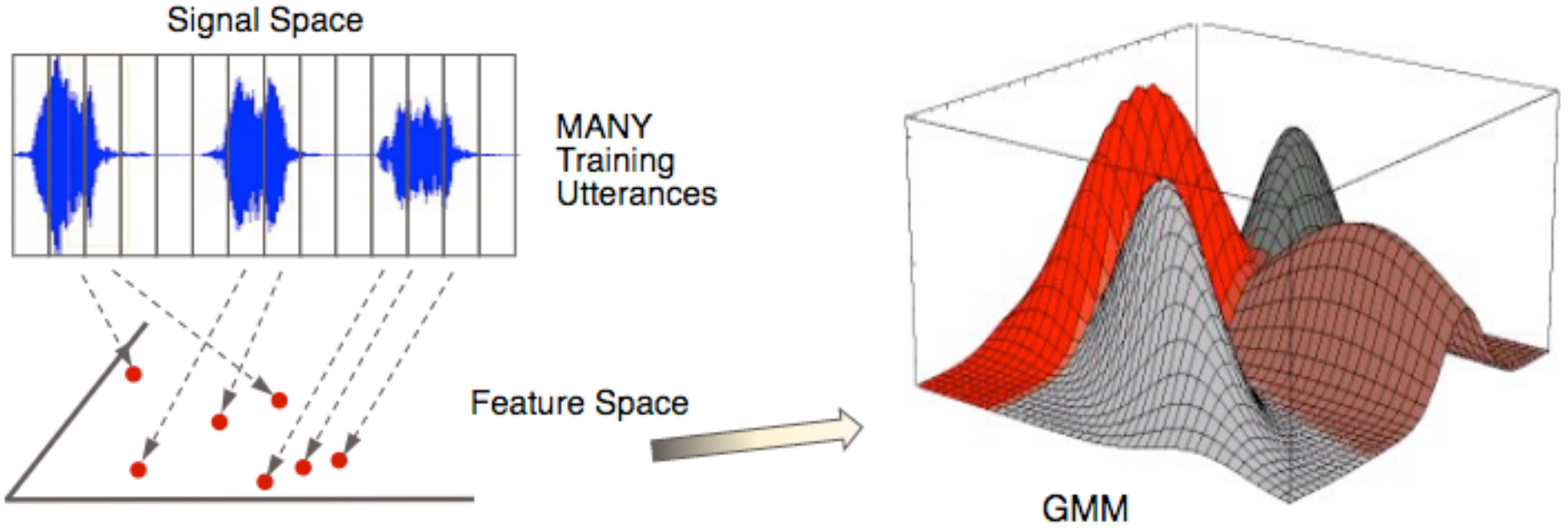
Extracting information from an utterance

- Most classification tasks have fixed input dimensions
  - If not, we divide into sections of fixed dimensions – frames
- But information about speakers comes from multiple frames

- How to extract a fixed dimensional information from varying input dimension?
  - Extract summary statistics
    - Build a model that summarize the utterance

Modeling speech

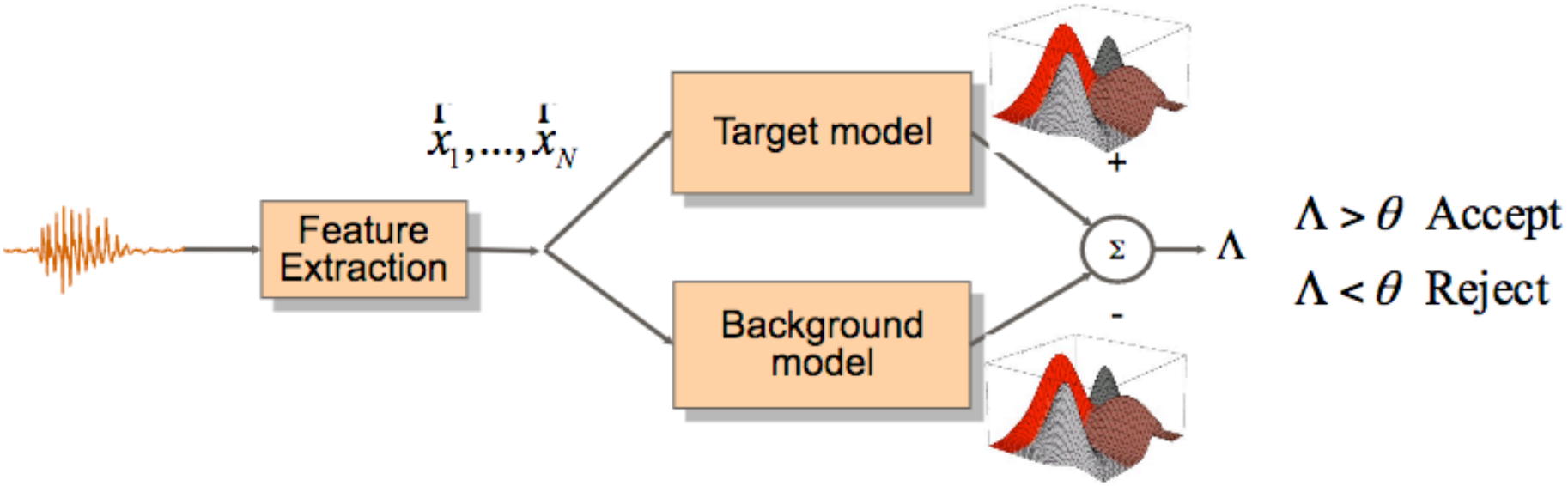
- We used GMMs for ASR
- Should be a good place to start!
- This GMM represents one class – speech from speaker X.
  - Each Gaussian in the mixture can represent each phonemes, etc. but we don't really care



A simple verification system

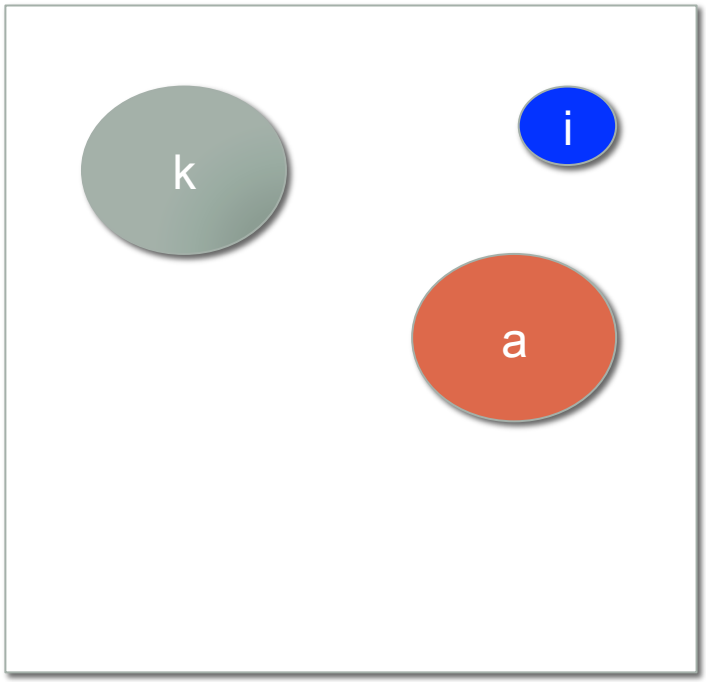
- Background model trained using multiple speakers
  - We call the background model Universal Background Model (UBM)
- To train the target model, we use enrollment sentences from the target speaker

$$LLR = \Lambda = \log p(X \mid \text{target}) - \log p(X \mid \overline{\text{target}})$$

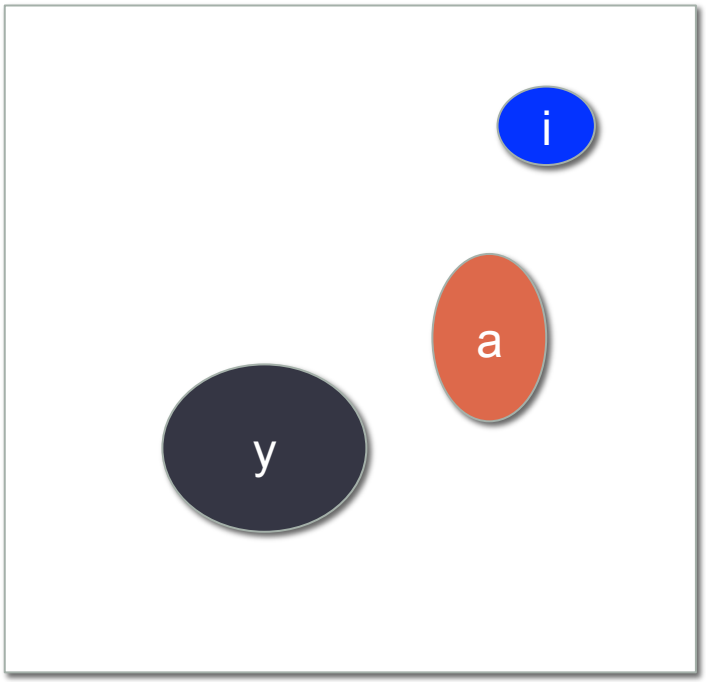


Not enough data

- Training a GMM requires many training data to be robust
- Requires lots of enrollment sentences, especially for text-independent speaker verification



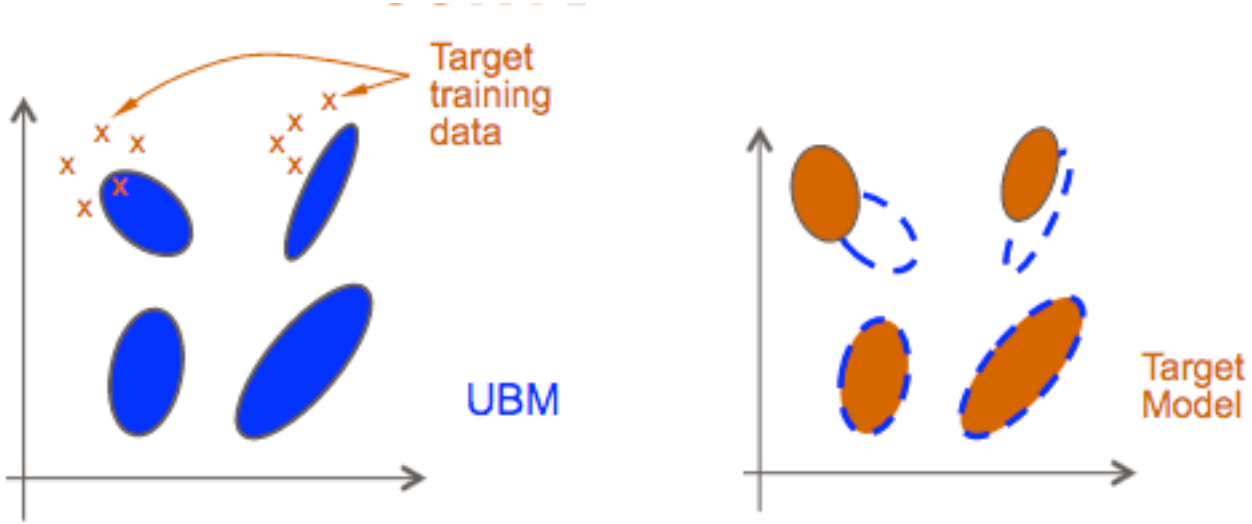
Model trained from sentence 1



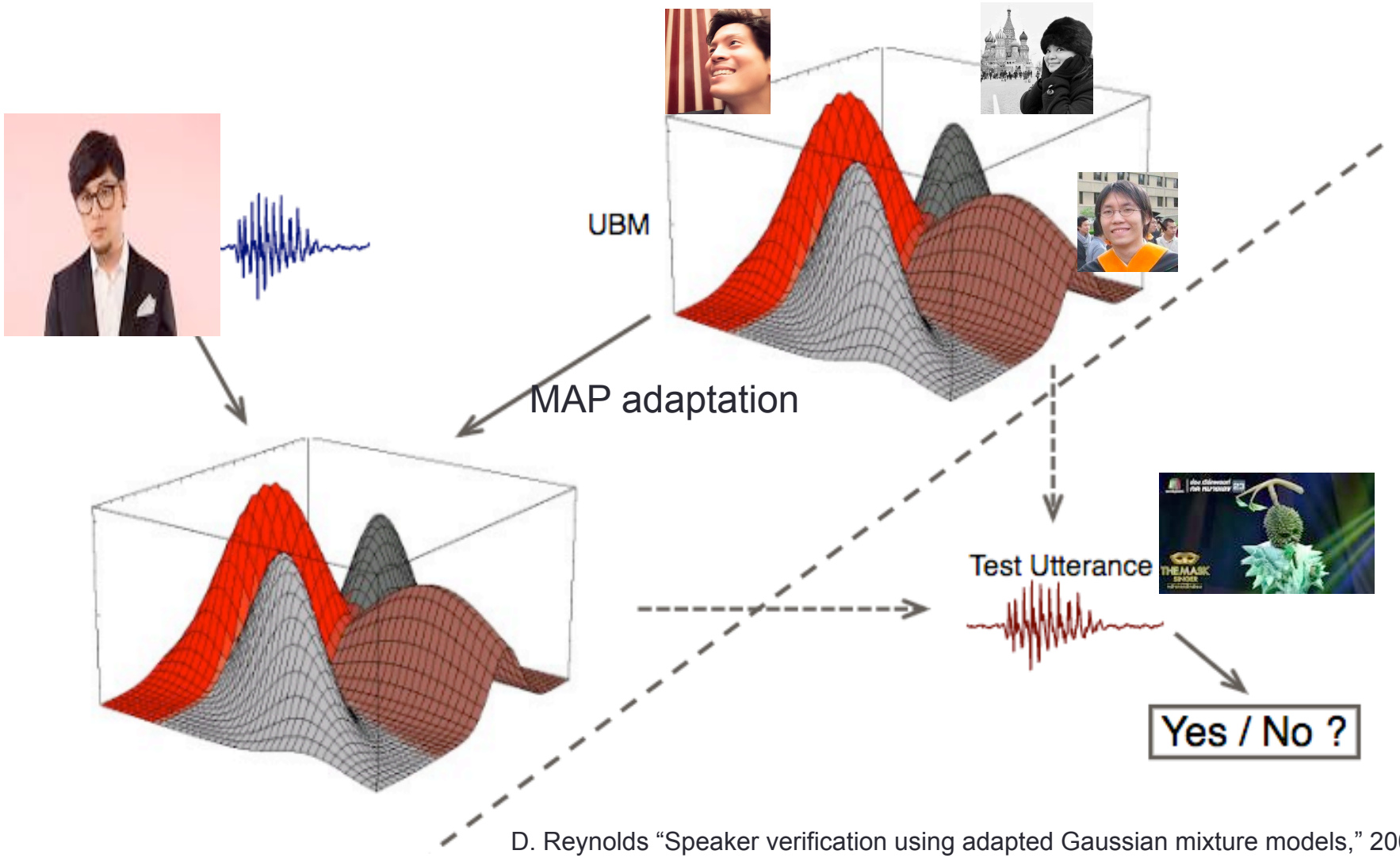
Model trained from sentence 2

Adaptation

- Adapt the target model from the UBM
- One way to adapt is Maximum A Priori (MAP) adaptation
  - Given adaptation frame
  - Compute which Gaussian it belongs to
  - Move the Guassian according to the new data (means only)
  - If little data, smaller movement. Large amount of data, bigger movement.



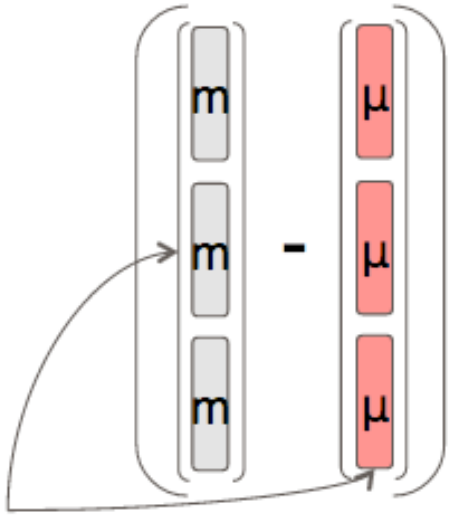
The GMM-UBM Approach



D. Reynolds “Speaker verification using adapted Gaussian mixture models,” 2000

Super-vector

- We can stack the means of the model into one long vector
- We assume this is enough to summarize the model since we only adapt the mean
- The difference summarizes how the enrollment data is special



Super-vectors

= special characteristic of enrollment data

MAP reformulated

- $M - m = Dz$
- $M = m + Dz$

where M is super-vector for the adapted GMM

m is the super-vector for the UBM

D is a diagonal matrix

z is the vector quantifying the enrollment data

- We assume z comes from a standard normal prior.
  - M is normally distributed with mean m and covariance D^2

A note on Dz

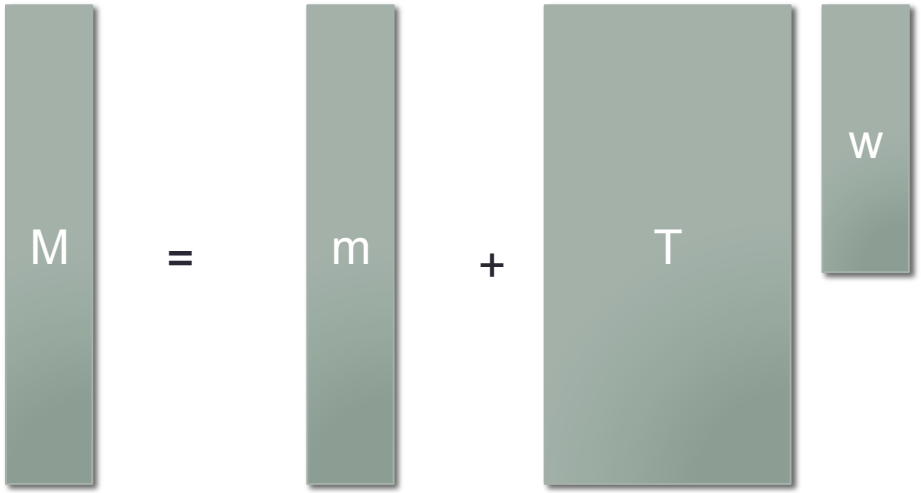
- The variability Dz captures the difference from the UBM in all aspects
  - Speaker
  - Channel (recording microphone)
  - Language
  - Emotion
  - Background noise, room acoustics
  - …
- z is very high dimensional, and from this formulation uncorrelated

Low dimensional subspace

- The movements of the mean should be constrained and correlated
  - Gaussian for /k/ and Gaussian for /g/ should move similarly
  - A male/female speech should move all the means in the same X direction

Total variability formulation

- M = m + Tw
  - T is a rectangular, low rank matrix (Total variability matrix)
  - w is a vector summarizing co-movements (Total factors, intermediate vector, i-vector)



Learning the T matrix

- Use Expectation Maximization (EM)
- Initialization: Pick a desired rank R for the Total Variability Matrix T and initialize this matrix randomly
- E step: For each utterance u, calculate the parameters of the posterior distribution of w(u) using the current estimates of m, T
- M step: Update T and by solving a set of linear equations using the w(u) from E step

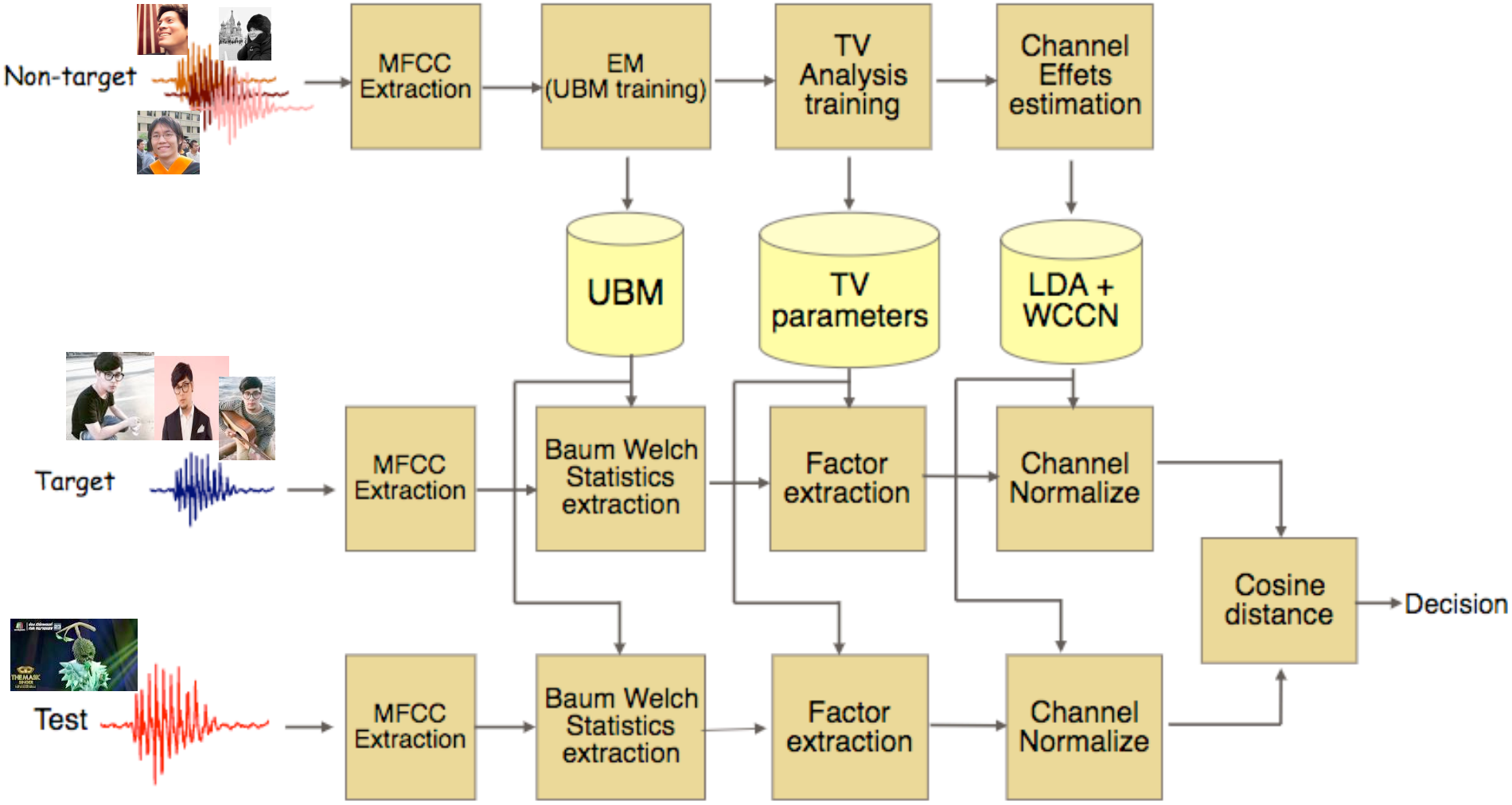
K. Boulianne, Eigenvoice Modeling with Sparse Training Data. IEEE Transactions on Speech and Audio Processing, 2005

- Or use PCA, works almost as well but easier to train

Picking the variability

- i-vector captures everything, speaker, channel, language, etc.
- Can we pick only the part we care about?
  - LDA – find the direction that differentiate between the class
  - Within-Class Covariance Normalization (WCCN) – normalize the covariance of each class

i-vector for speaker verification



N. Dehak Front-end factor analysis for speaker verification, 2011

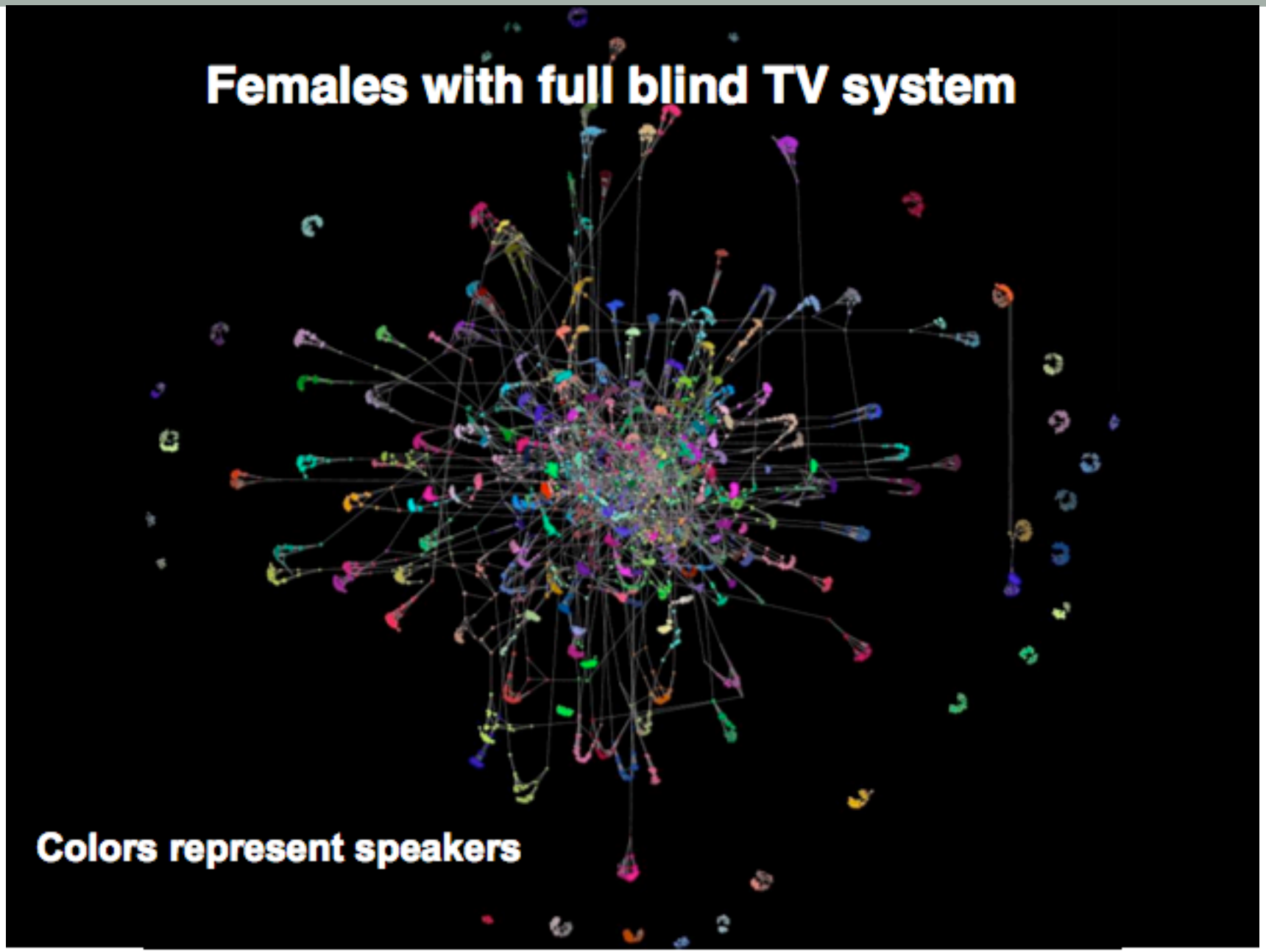
Speaker recognition performance

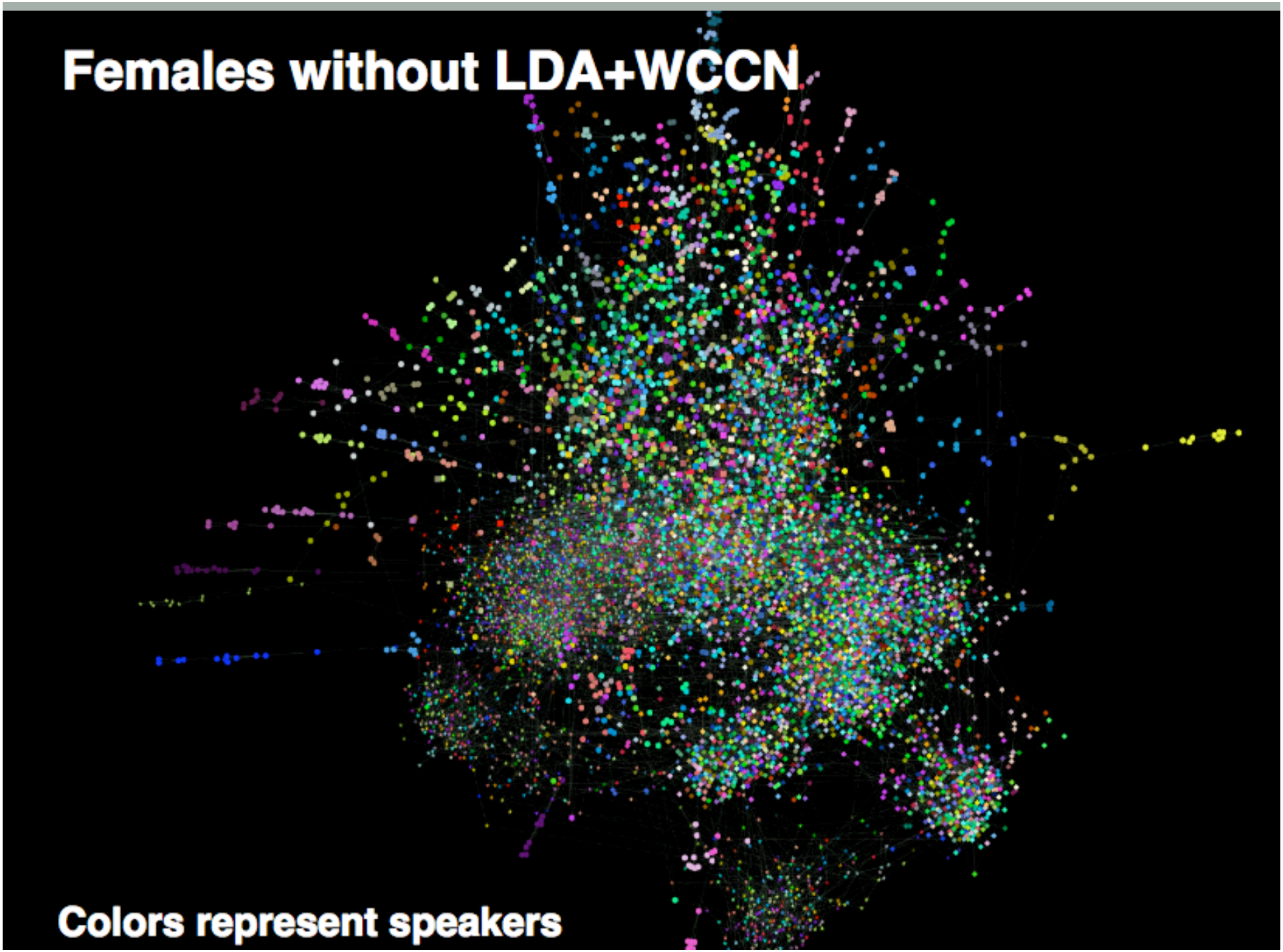
- Speaker Recognition Evaluation 2008
  - ~1000 female ~600 male speakers
  - Contain telephone and microphone recordings
  - 40k recordings
  - Tasks:
    - 10 sec enrollment –10 sec test
    - 1 conversation enrollment (~2 mins) – 1 conversation test

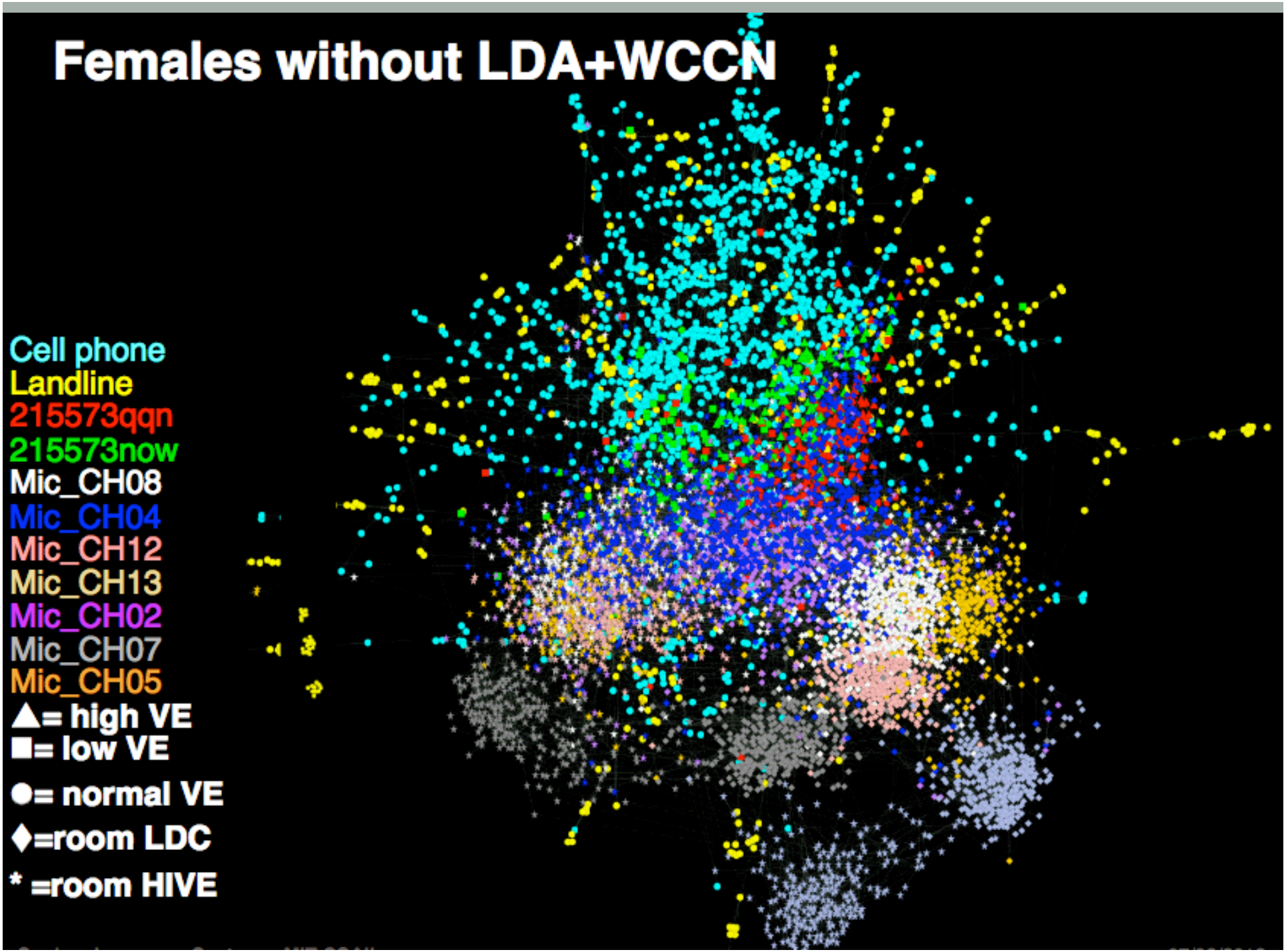
Speaker recognition performance

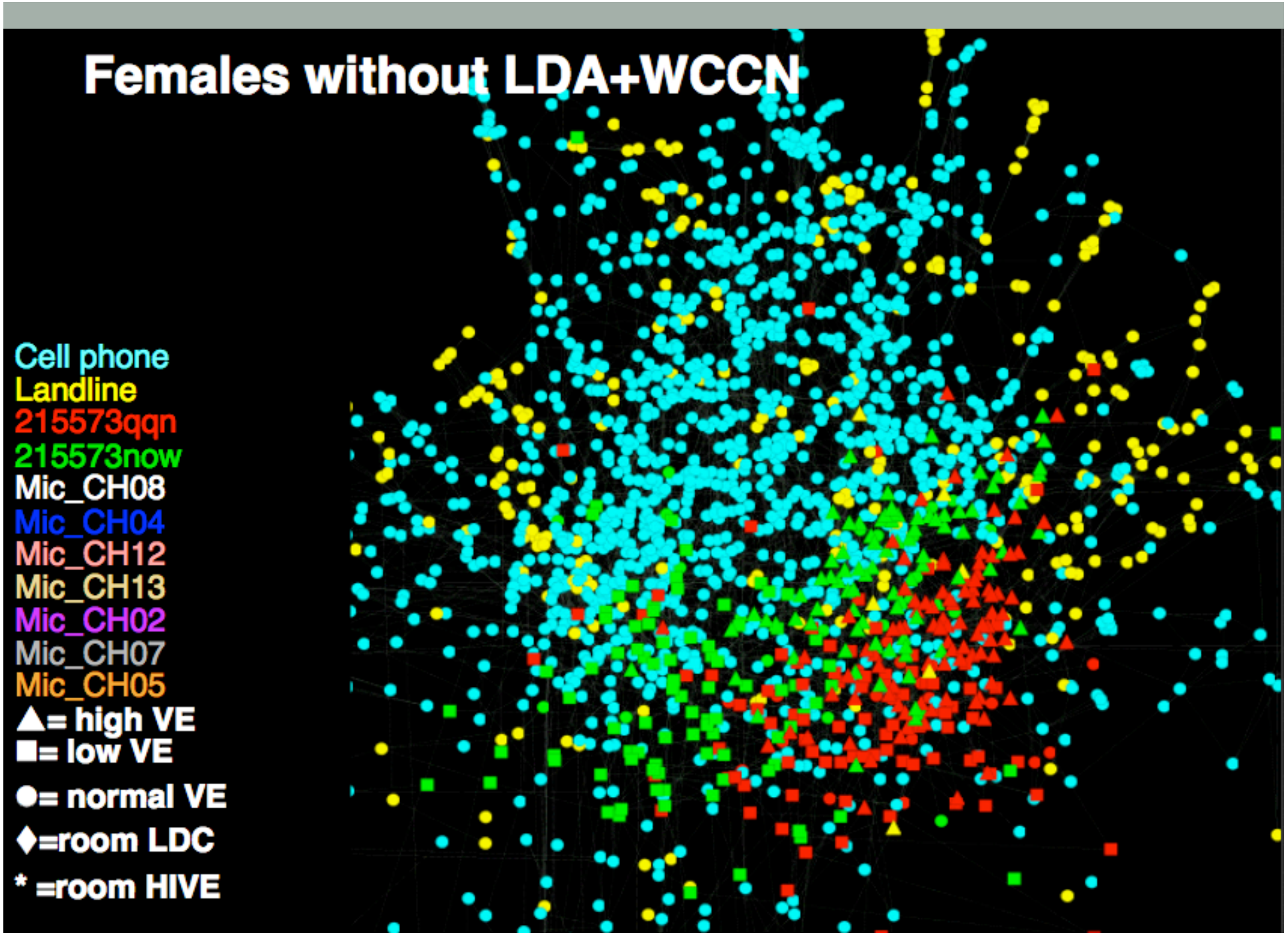
	English only EER	All language EER
Female conv-conv	2.9%	5.76%
Male conv-conv	1.12%	4.48%
Female 10-10	12.19%	16.59%
Male 10-10	11.09%	14.44%

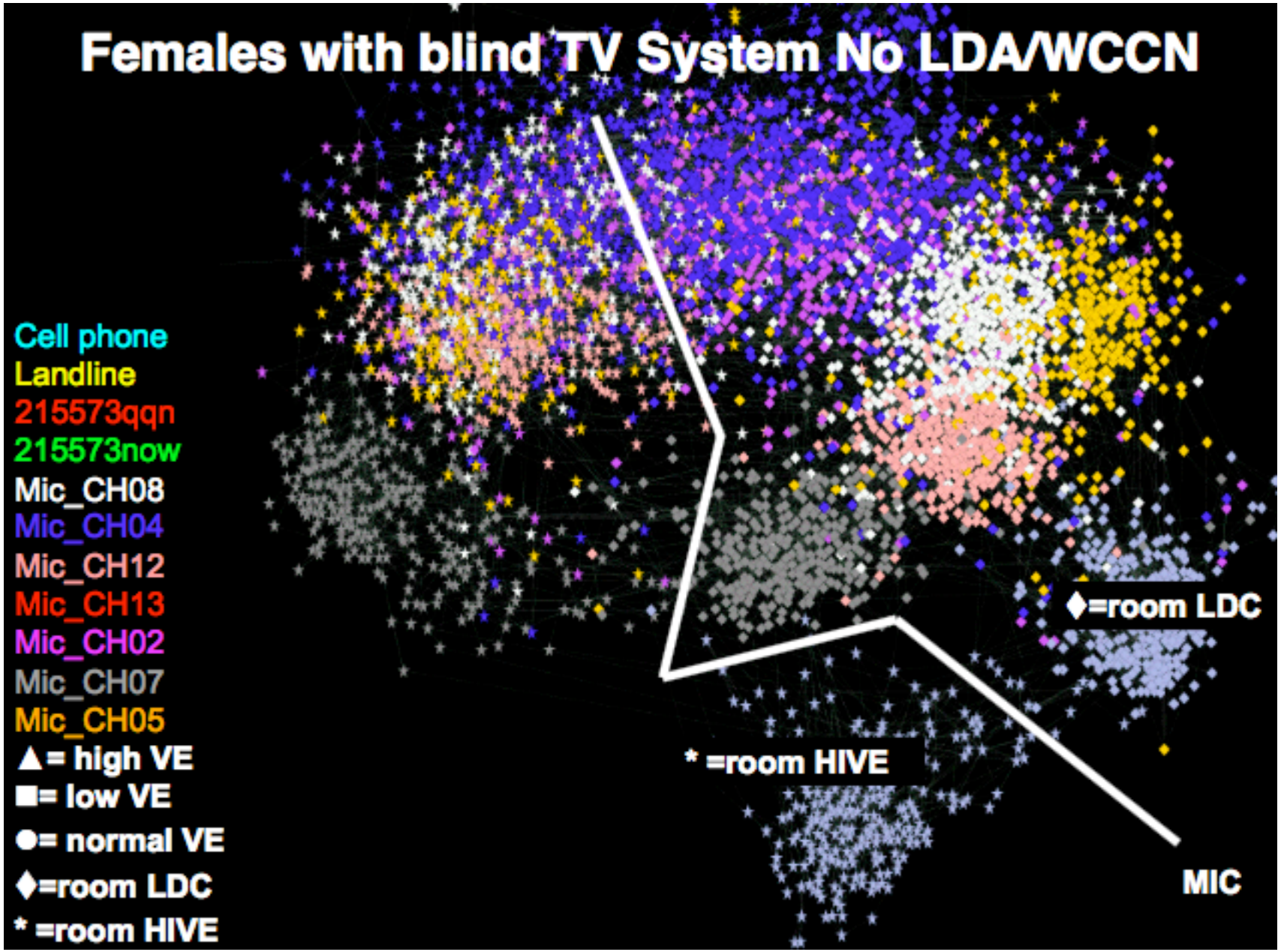
Performance is better than human listener!

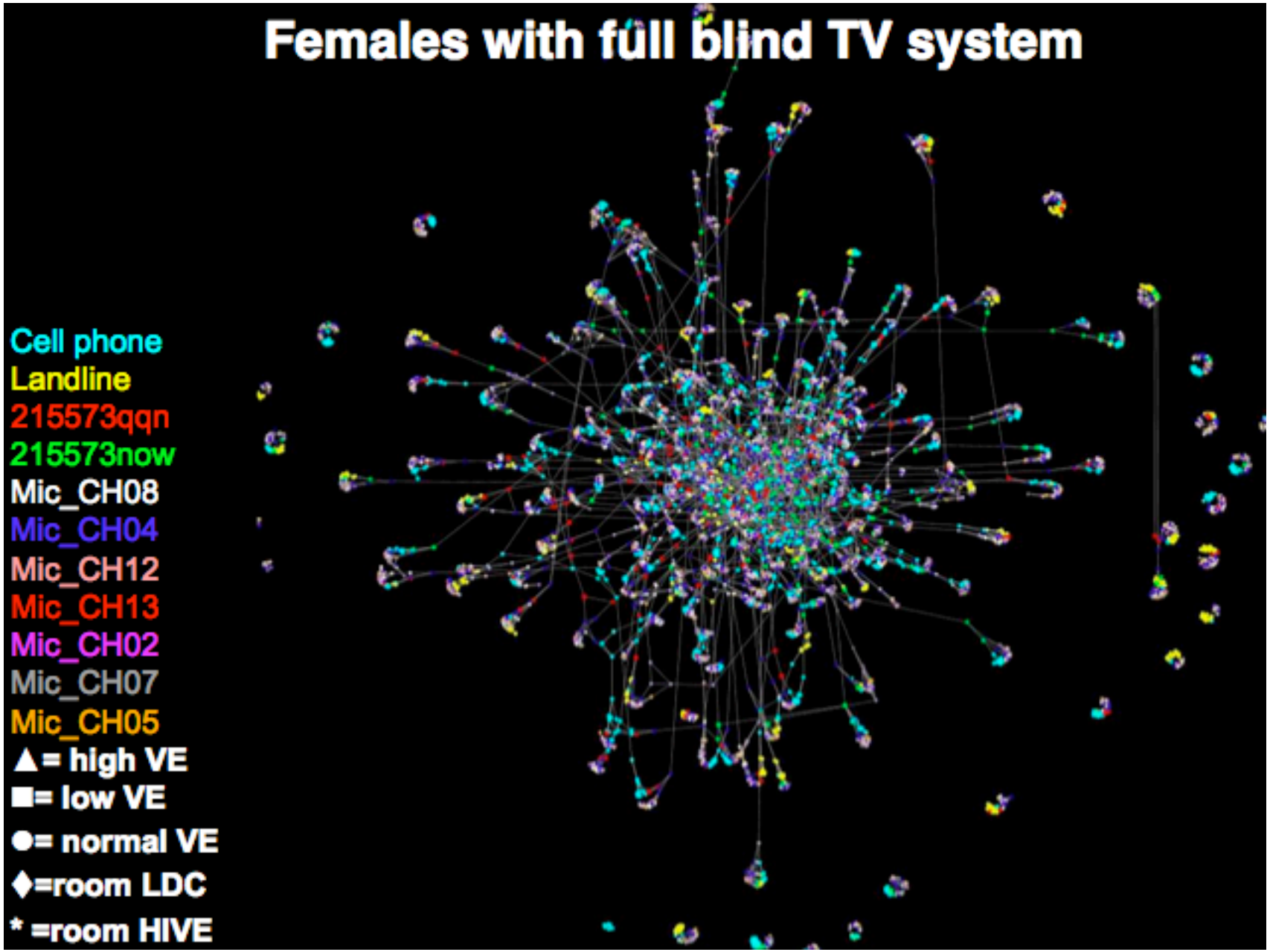

Females with full blind TV system
Colors represent speakers

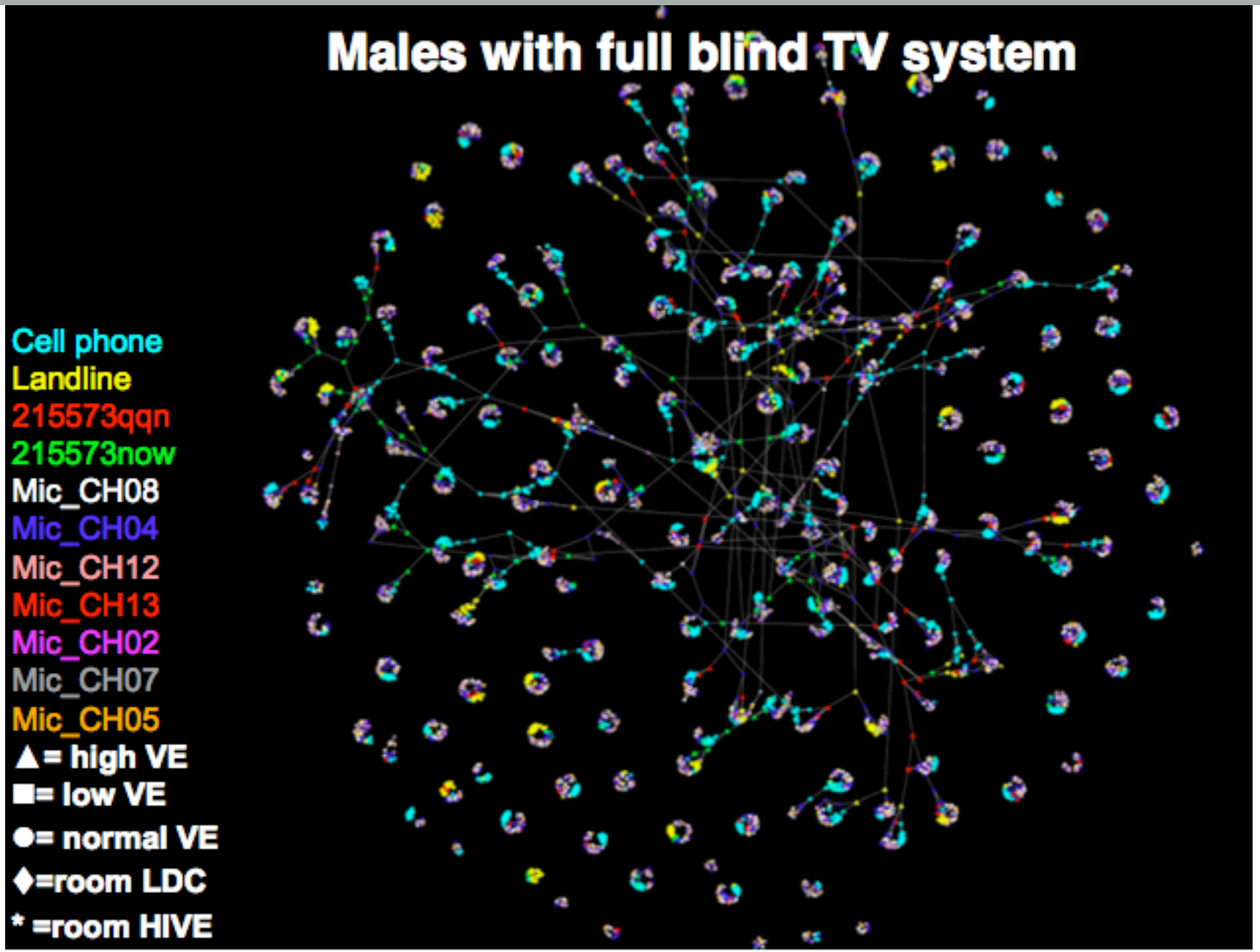

Females without LDA+WCCN
Colors represent speakers


Females without LDA+WCCN
Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE


Females without LDA+WCCN
Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE


Females with blind TV System No LDA/WCCN
Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE
♦=room LDC
* =room HIVE
MIC


Females with full blind TV system
Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE


Males with full blind TV system
Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE

Current trend in speaker-ID

- Move to harder tasks: harder channel condition, room acoustics
  - Using less data, or out of domain data
- DNN+i-vector give some performance boost
- Detecting recorded or synthesized speech

Speaker ID summary

- Summarize utterance statistics using means of a GMM
- Estimate the movements of the means in a lower dimensional subspace
- Extract only the information we need using LDA

Today topics

- VAD
- Noise reduction
  - Automatic Gain control
  - spectral subtraction
  - adaptive filtering
  - microphone array
  - Multi-condition training
- Language ID/Speaker ID/Emotion ID
  - i-vector
- Adaptation
  - VTLN
  - fMLLR
  - DNN adaptation
- Semi-supervised training and crowd sourcing
- Keyword search

Speaker adaptation

- What we used so far are speaker independent model
  - Same model for all speakers
- We can also adapt the model according to the target speaker
  - More data for adaptation gives better effect
- Usually requires multiple decoding passes
  - First pass to get the speaker estimate
  - Then use the estimate to adapt the model for final decode

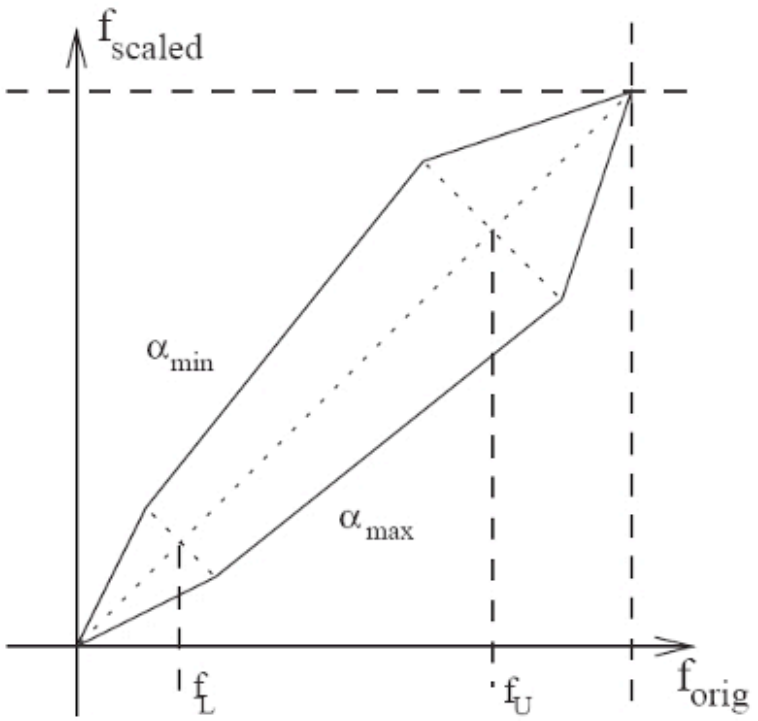
Speaker adaptation

Two main approach for adaptation

- Feature domain adaption
  - Find a transform that makes the feature fit more to the model
- Model domain adaptation
  - MAP adaptation

Vocal Tract Length Normalization (VTLN)

- Vocal tract length effects formant frequencies
  - Longer vocal tracts give lower formant frequencies
- Goal: estimate the speaker vocal tract length and normalize the formant frequencies
- Model the effect of vocal tract length as frequency shifts
- Shifts the Mel filterbank using the scaled frequencies
- Estimate the warping factor for each speaker using EM



VTLN results

Table 1: Results of recognition experiments (%)

	Male	Female	Ave.
SI	78.7	78.8	78.8
ML-VTLN	79.4	79.1	79.3

- Small gain in performance
- Can be used for data augmentation (see data augmentation sides)

T. Emori, Vocal Tract Length Normalization using Maximum Likelihood Estimation, 2001

Feature space Maximum Likelihood Linear Regression (fMLLR)

- Goal: find a projection matrix that maximize the likelihood for the target speaker
- Most popular adaptation technique for HMM-GMM
- Can be used to adapt the features for DNN system
- Usually give around 10% relative performance improvement

WER on LVCSR Mandarin Task

None	fMLLR
20.4%	17.9%

D. Povey, Feature and model space speaker adaptation with full covariance Gaussians 2006

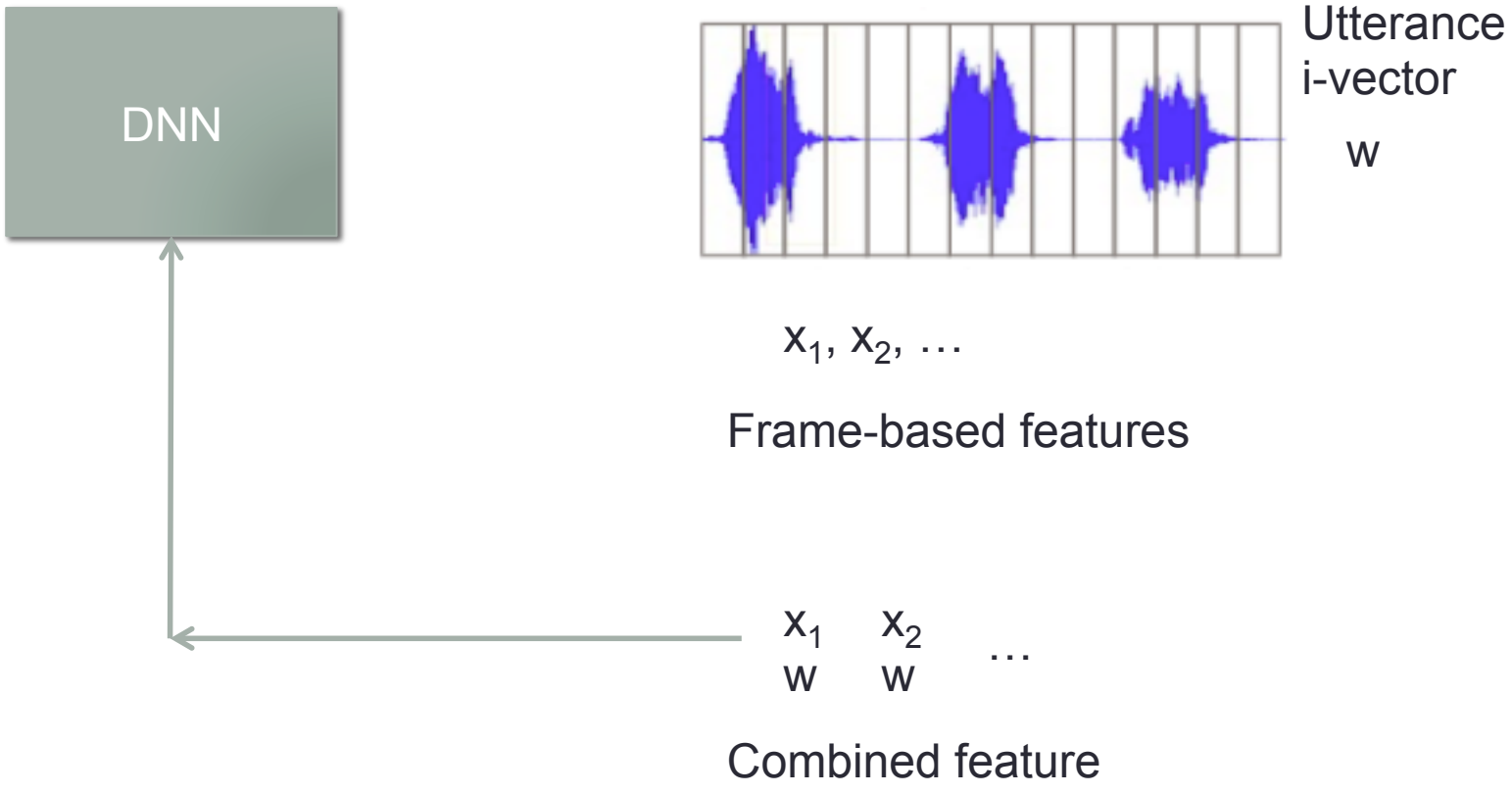
DNN adaptation

- The previous methods can perform adaptation even on small amount of adaptation data (<10 sentences)
- Cannot adapt DNN (using SGD) with such tiny amount of data

- Give the DNN auxiliary information to help with ASR instead
  - Is the speaker male/female?
  - Recording microphone
  - Noise information
  - i-vector

i-vector-based DNN adaptation

- Append i-vector to normal input features



i-vector adaptation results

- Note this is not really speaker adaptation but utterance adaptation
- i-vector adaptation is included in your project AM

WER on voice search

Model size	i-vector dimensions (k)					
	0	20	50	100	200	300
Small	17.8	**17.0**	17.2	17.4	17.9	18.2
Medium	15.0	**14.5**	**14.5**	**14.5**	15.2	15.5
Large	11.0	**10.9**	**10.9**	11.2	11.8	12.3

- DNN seems to like this auxiliary info
  - Bottleneck feature extracted from a single random frame of a video helps ASR of the entire video

Andrew Senior, Improving DNN Speaker Independence with I-vector Inputs, 2014

Today topics

- VAD
- Noise reduction
  - Automatic Gain control
  - spectral subtraction
  - adaptive filtering
  - microphone array
  - Multi-condition training
- Language ID/Speaker ID/Emotion ID
  - i-vector
- Adaptation
  - VTLN
  - fMLLR
  - DNN adaptation
- Semi-supervised training and crowd sourcing
- Keyword search

Semi-supervised training

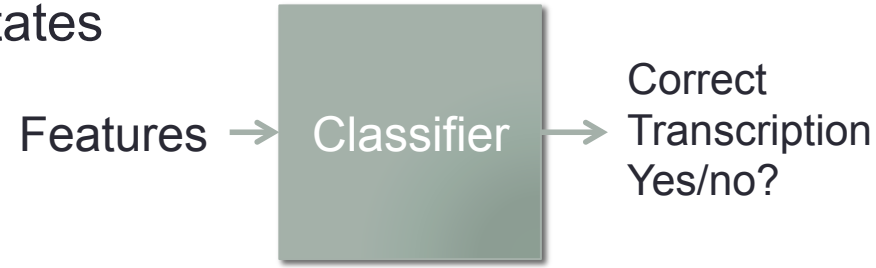
- Training from labeled data is supervised training
- Training from unlabeled data is unsupervised training
- Semi-supervised has both the supervised and unsupervised portion
  - Train a model on a small labeled data set
  - Use the model to classify and label a bigger data set
  - Treat the classification results are correct and train a semi-supervised system

Semi-supervised training

- Care need to be used when using semi-supervised training
  - ASR has errors
  - Using bad transcriptions hurts performance
  - Using perfect transcriptions does not help the recognizer identify its weakness
- Needs a way to quantify confidence of a recognition result
  - Cannot directly use probability from the model – longer sentence has lower probability than shorter one

Confidence measures

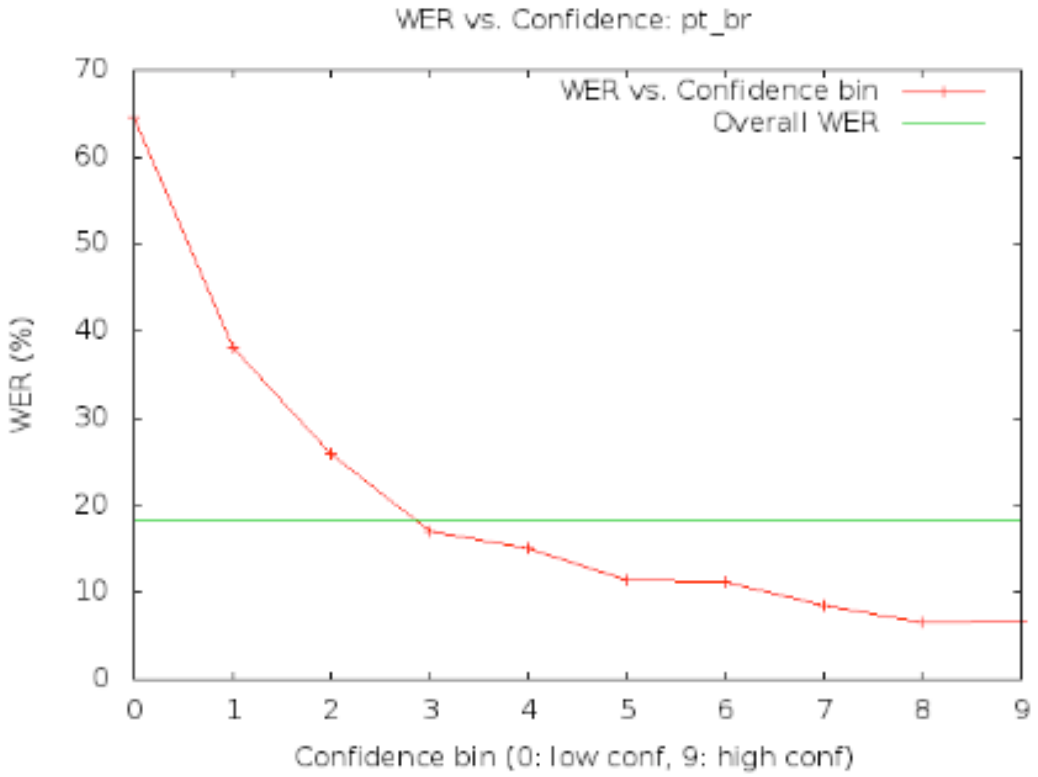
- Create another classifier to estimate confidence of correctness
- Features
  - Statistics about the log likelihood score
  - N-best characteristics
  - Duration about phonemes, hmm states
  - LM AM scores
  - Posterior probabilities
  - Etc.



- Use a held out set to train the classifier to predict whether a hypothesis is correct or not
  - A simple logistic regression works

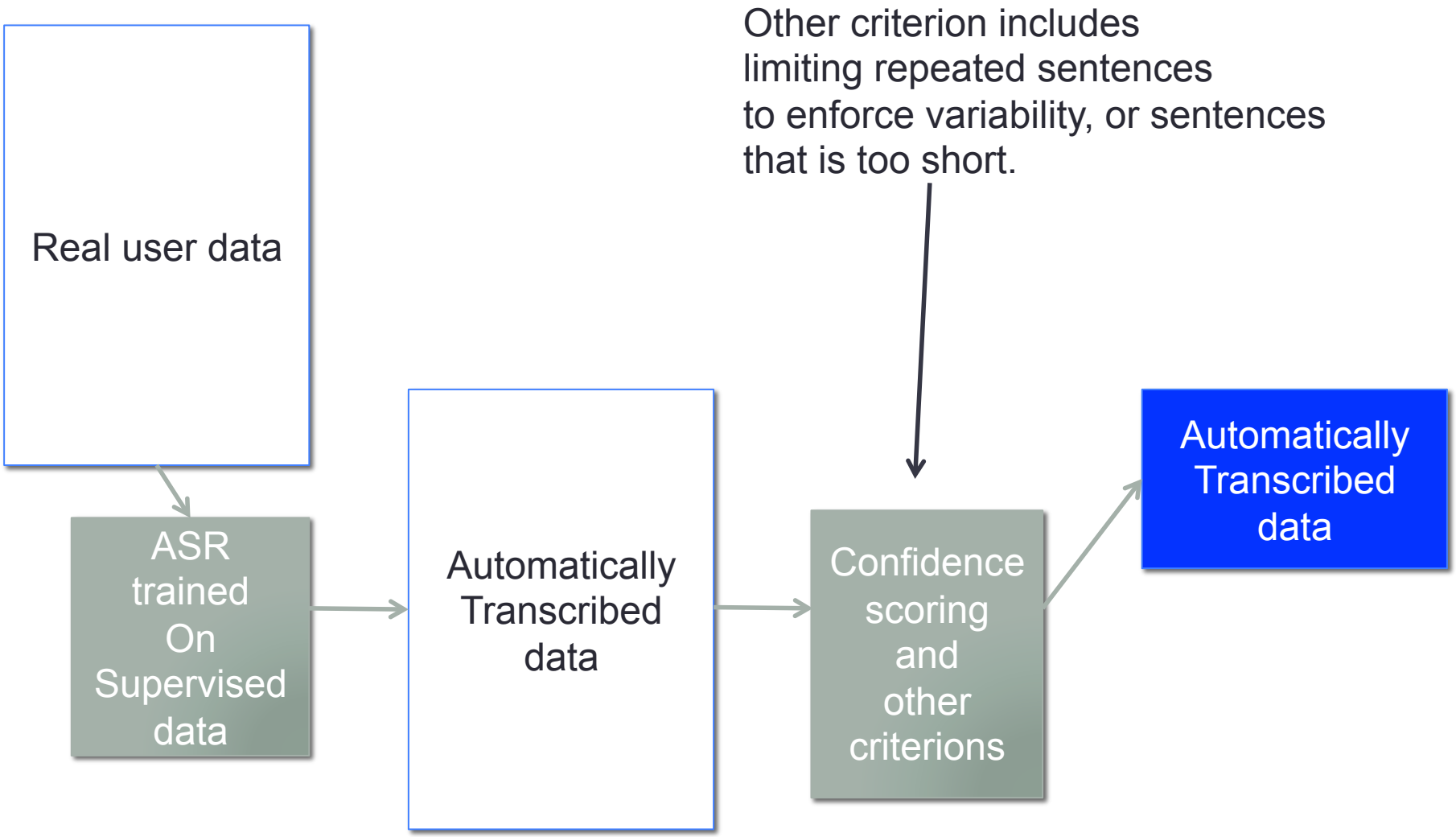
Confidence vs WER

- Having accurate confidence score is hard, but it still give meaningful information



O. Kapralova, A big data approach to acoustic model training corpus selection 2014

Semi-supervised training framework



Semi-supervised results

Language	WER baseline	WER proposed	Relative gain
Russian	27.5	25.1	8.7%
French	16.2	14.6	10.4%
Italian	13.6	12.1	11%
Brazilian Portuguese	24.3	20.9	14%

Table 4: Performance comparison of the baseline approach (2M training set, production transcripts) and the proposed approach (20M training set, redecoded transcripts)

O. Kapralova, A big data approach to acoustic model training corpus selection 2014

Crowd-sourcing

- What is crowd-sourcing?

Example of data collection with crowd

Crowd-collection: Addresses

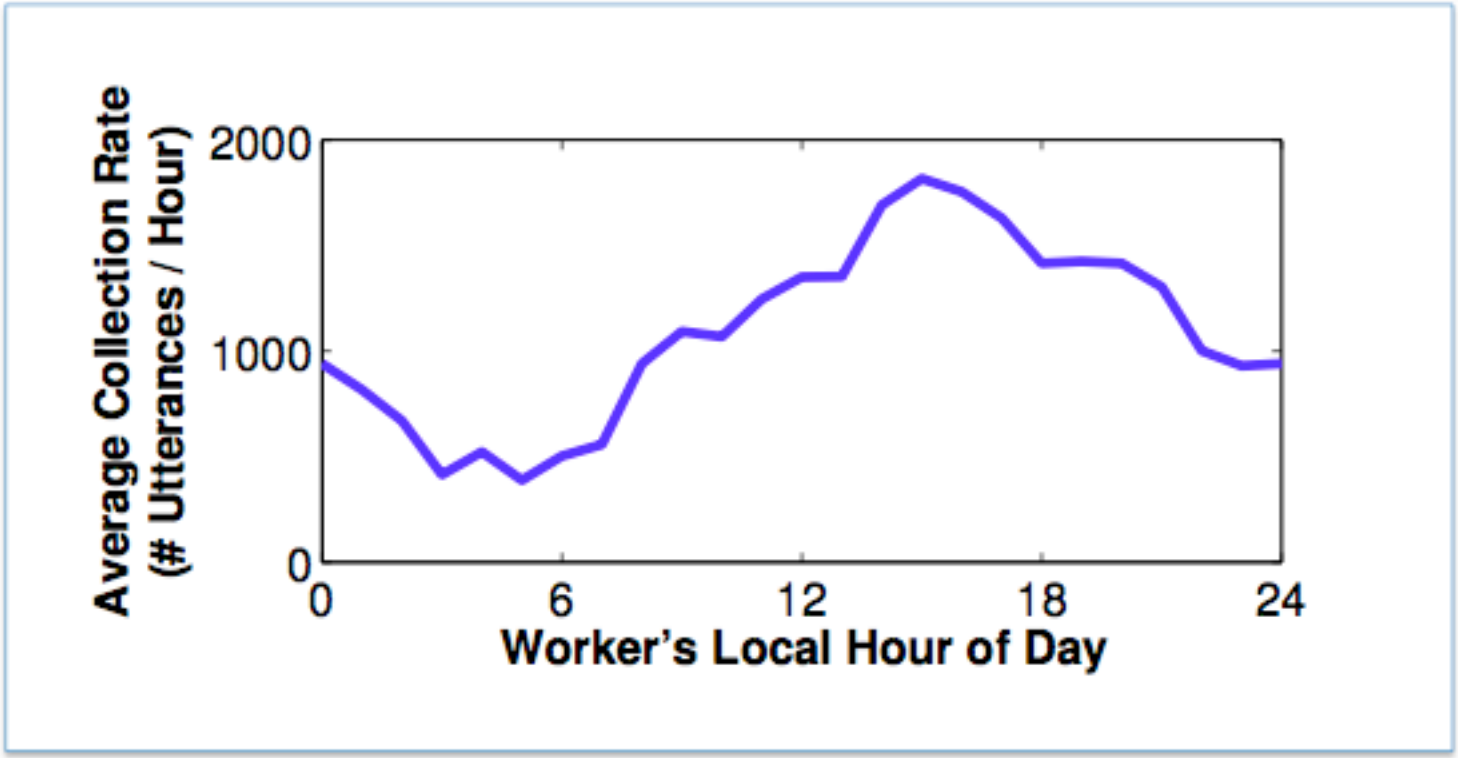
100,000 prompts

1¢ a piece.

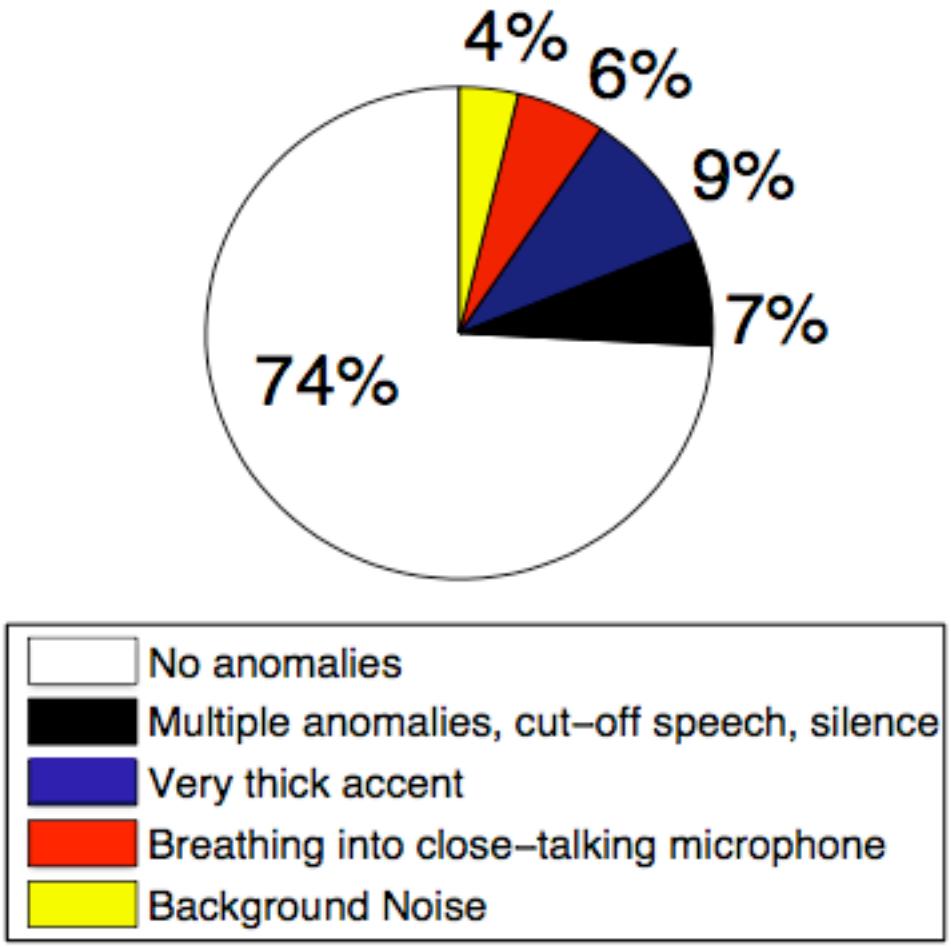
I. McGraw, Crowd-supervised Training of Spoken Language Systems, 2012

Example of data collection with crowd

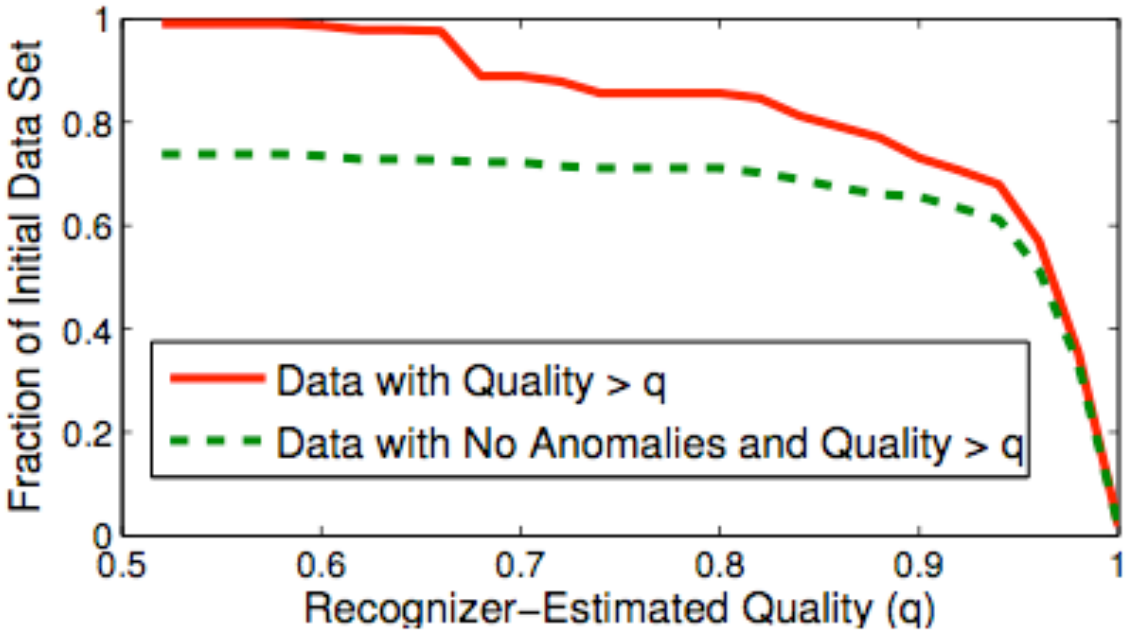
Collected 103 hours of audio in 77 hours.



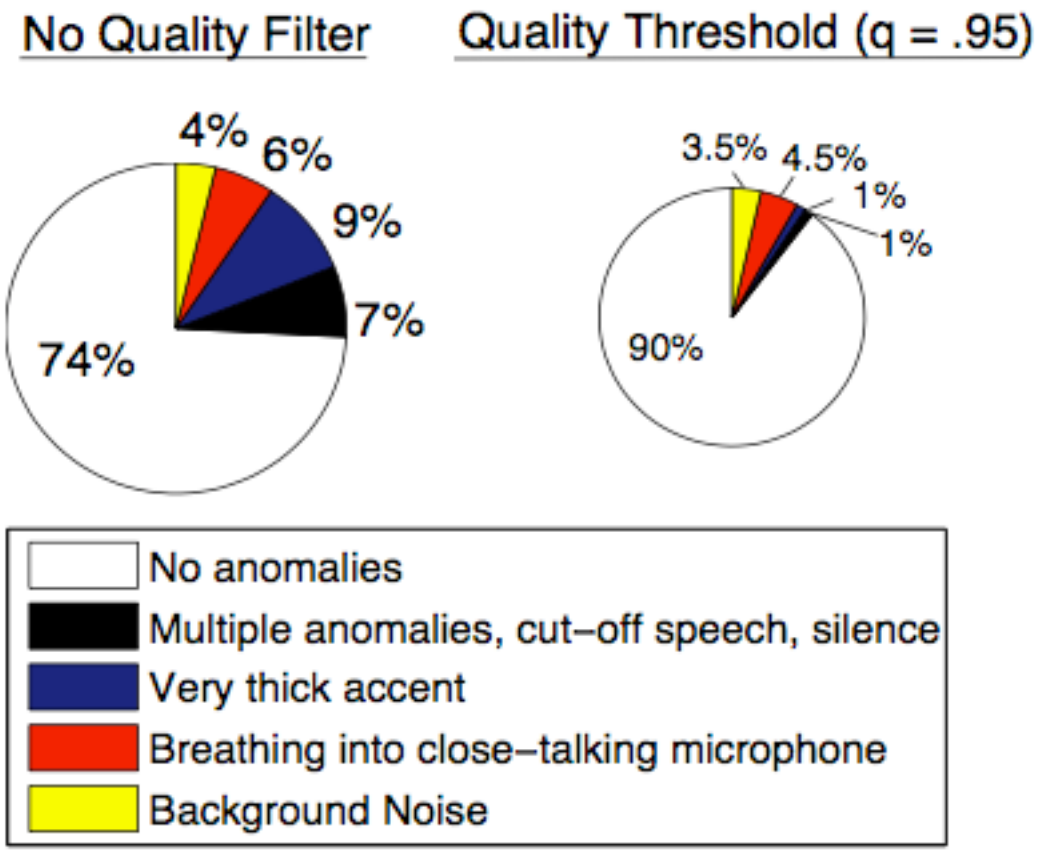
Speech Characteristics



Filtering using confidence



Quality after filtering

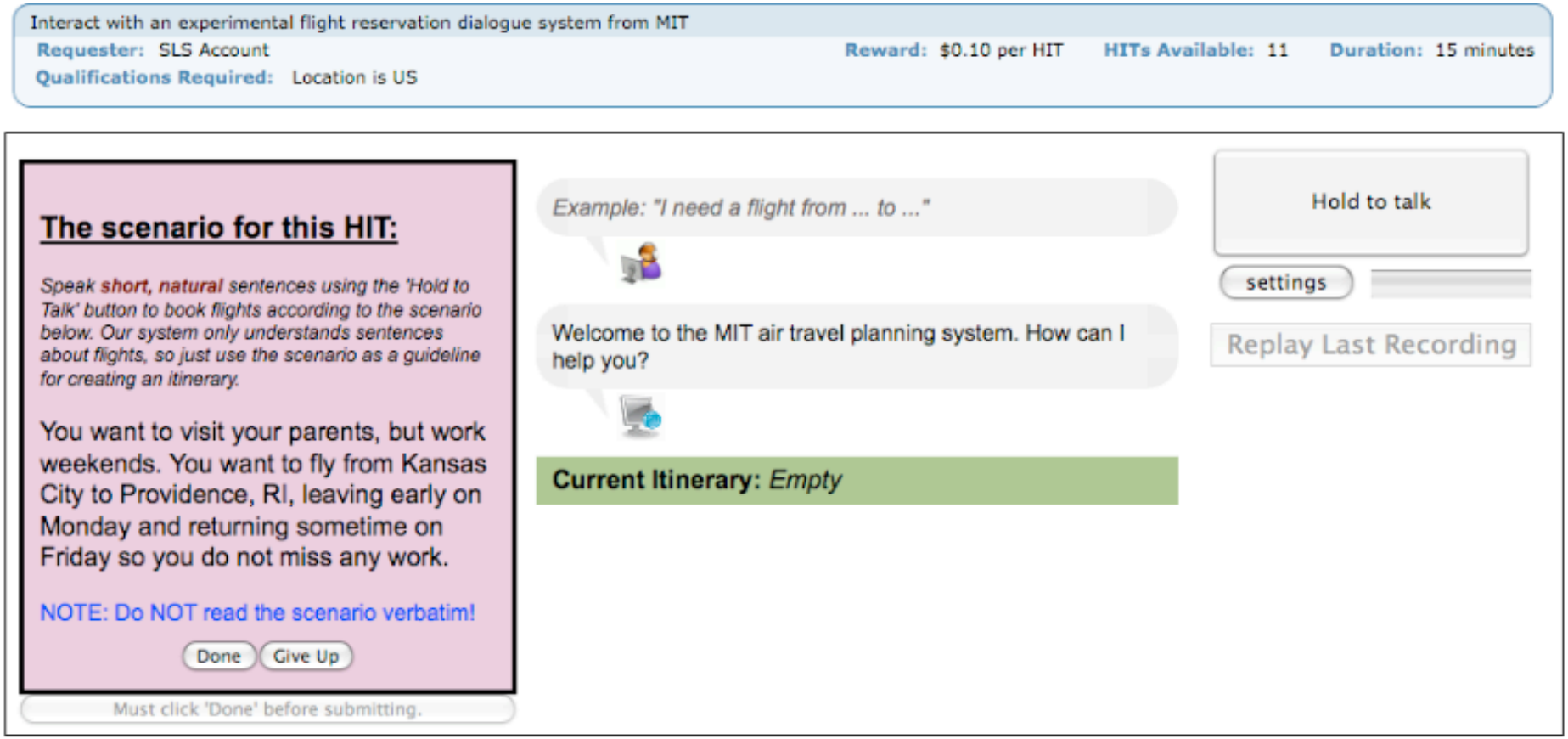


Filtering errors

- Need a way to handle errors/cheaters
  - Have a task that you know the answer
  - Use ASR to verify
  - Filter by work rating
  - Voting
  - Manual verification

Other usage

- Transcription
- Coming up with example use scenarios
- Coming up with example sentences
- Test your system
- Semantic tagging

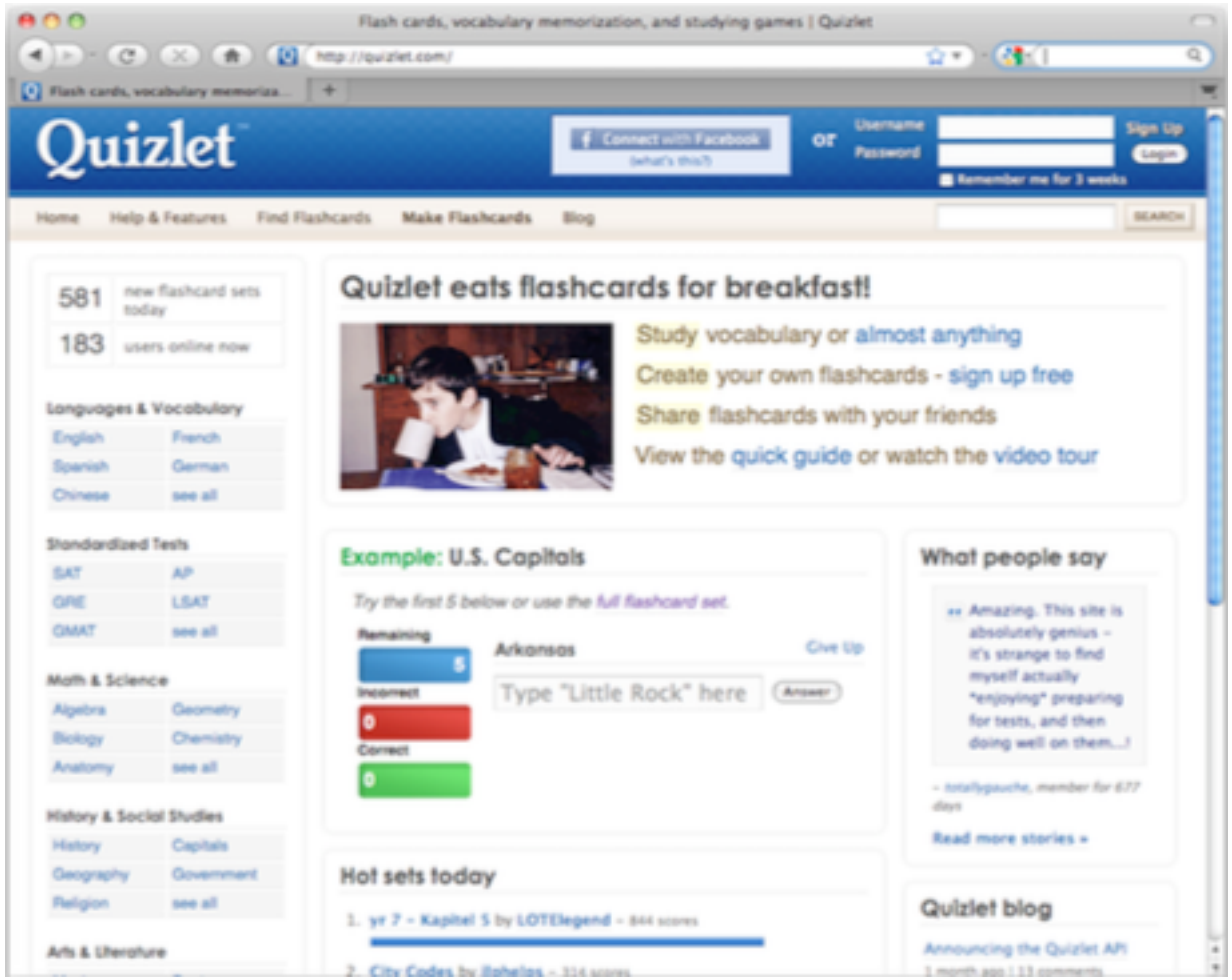

Interact with an experimental flight reservation dialogue system from MIT
Requester: SLS Account
Reward: $0.10 per HIT
HITs Available: 11
Duration: 15 minutes
Qualifications Required: Location is US
The scenario for this HIT:
Speak short, natural sentences using the 'Hold to Talk' button to book flights according to the scenario below. Our system only understands sentences about flights, so just use the scenario as a guideline for creating an itinerary.
You want to visit your parents, but work weekends. You want to fly from Kansas City to Providence, RI, leaving early on Monday and returning sometime on Friday so you do not miss any work.
NOTE: Do NOT read the scenario verbatim!
Done
Give Up
Must click 'Done' before submitting.
Example: "I need a flight from ... to ..."
Welcome to the MIT air travel planning system. How can I help you?
Current Itinerary: Empty
Hold to talk
settings
Replay Last Recording

Another way to use crowd info

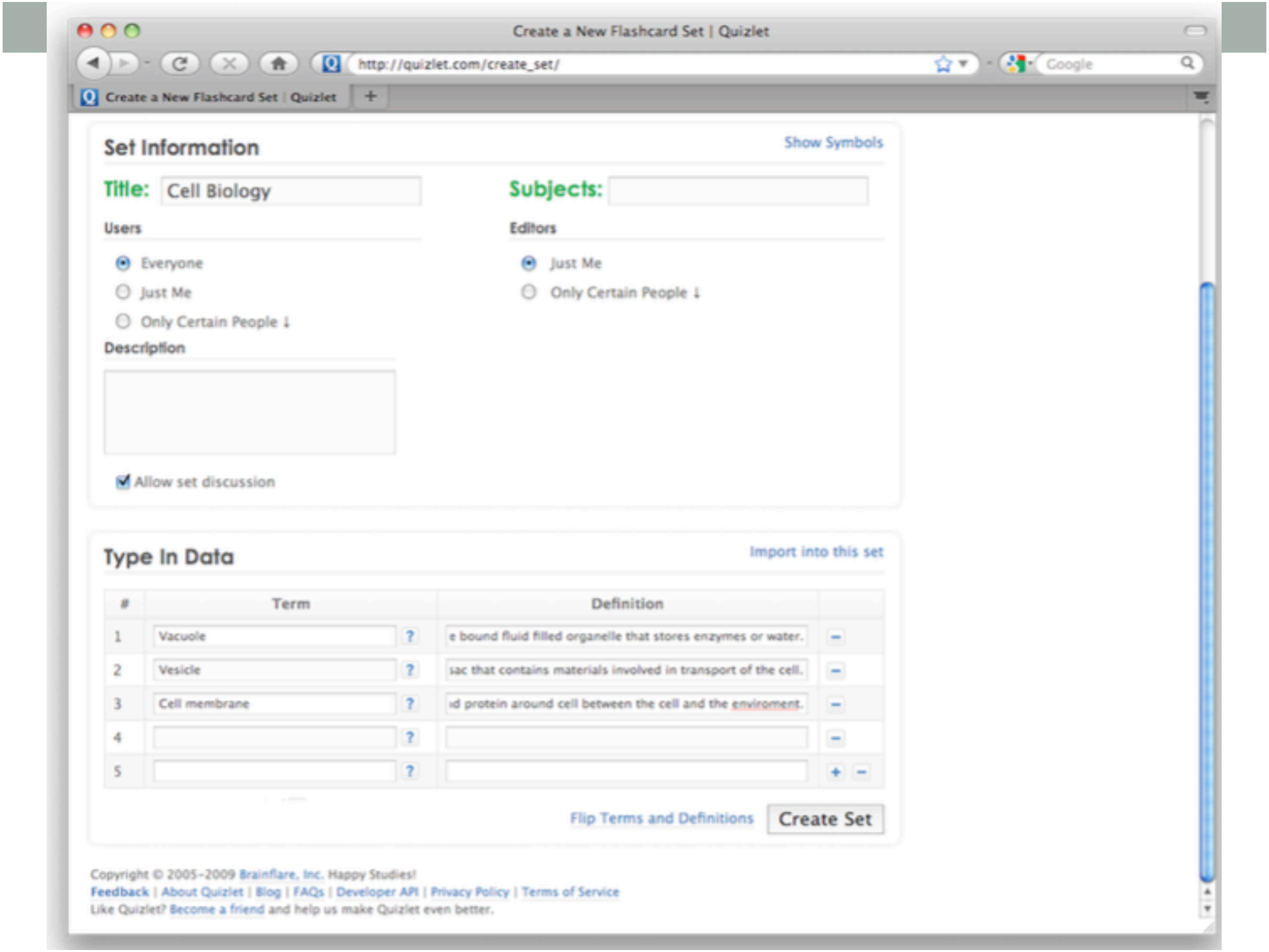
- User behaviors can give powerful information
  - Google search – clicking on search results

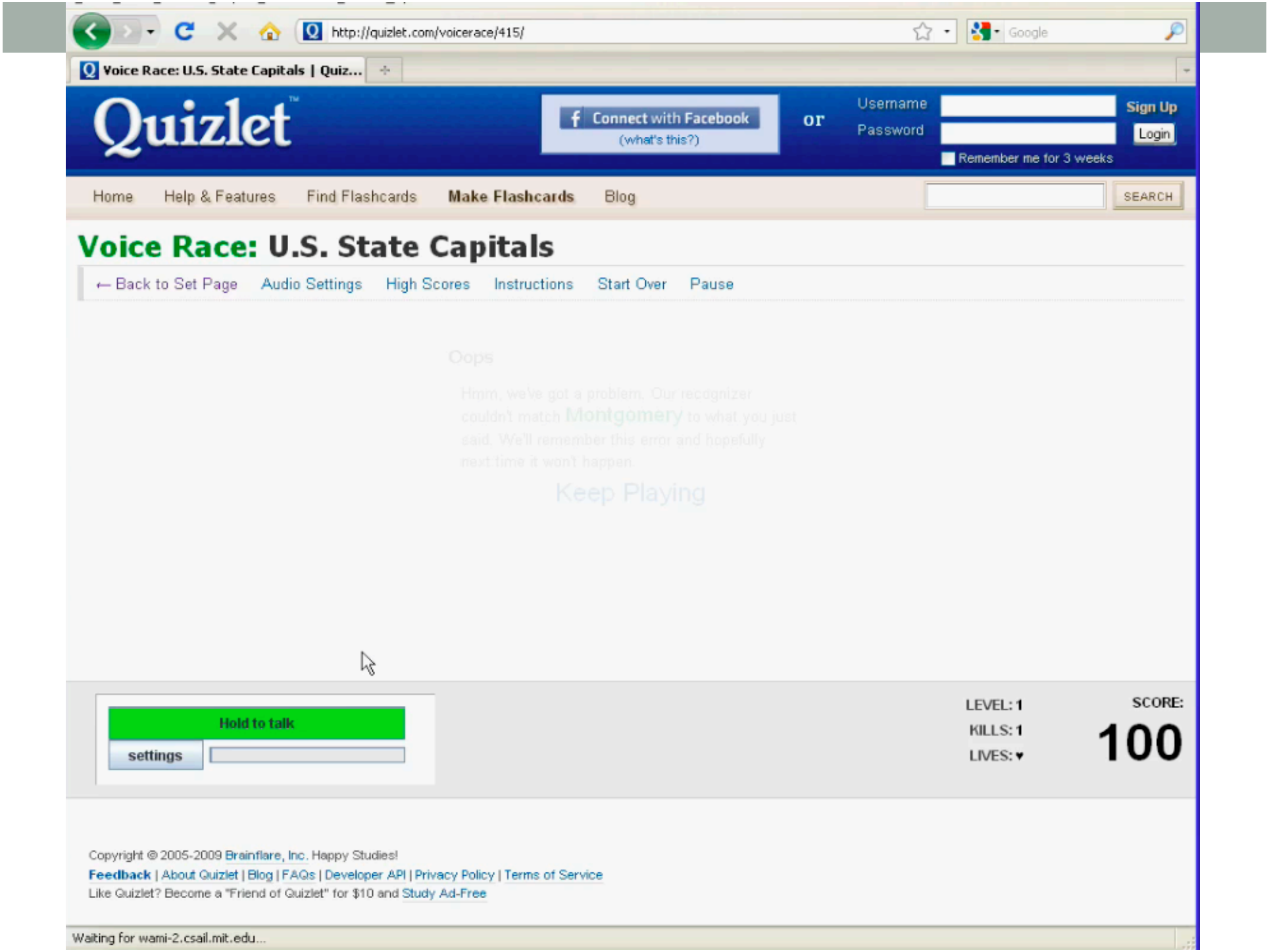
- No monetary incentive required, but need to structure the task so that the information is useful

Self-supervised speech interface



I. McGraw, A Self-Labeling Speech Corpus: Collecting Spoken Words with an Online Educational Game, 2009.


Create a New Flashcard Set | Quizlet
http://quizlet.com/create_set/
Google
Create a New Flashcard Set | Quizlet
Set Information
Show Symbols
Title: Cell Biology
Subjects:
Users
Everyone
Just Me
Only Certain People ↓
Editors
Just Me
Only Certain People ↓
Description
Allow set discussion
Type In Data
Import into this set
Term Definition
1 Vacuole e bound fluid filled organelle that stores enzymes or water.
2 Vesicle sac that contains materials involved in transport of the cell.
3 Cell membrane id protein around cell between the cell and the enviroment.
4
5
Flip Terms and Definitions
Create Set
Copyright © 2005-2009 Brainflare, Inc. Happy Studies!
Feedback | About Quizlet | Blog | FAQs | Developer API | Privacy Policy | Terms of Service
Like Quizlet? Become a friend and help us make Quizlet even better.


http://quizlet.com/voicerace/415/
Google
Voice Race: U.S. State Capitals | Quiz...
Quizlet
Connect with Facebook
(what's this?)
or
Username
Password
Sign Up
Login
Remember me for 3 weeks
Home
Help & Features
Find Flashcards
Make Flashcards
Blog
SEARCH
Voice Race: U.S. State Capitals
← Back to Set Page
Audio Settings
High Scores
Instructions
Start Over
Pause
Oops
Hmm, we've got a problem. Our recognizer couldn't match Montgomery to what you just said. We'll remember this error and hopefully next time it won't happen.
Keep Playing
Hold to talk
settings
LEVEL: 1
KILLS: 1
LIVES: ♥
SCORE:
100
Copyright © 2005-2009 Brainflare, Inc. Happy Studies!
Feedback | About Quizlet | Blog | FAQs | Developer API | Privacy Policy | Terms of Service
Like Quizlet? Become a "Friend of Quizlet" for $10 and Study Ad-Free
Waiting for wami-2.csail.mit.edu...

ASR setup

- Deployed for 22 days
- Vocab is just the words in the quiz
- LM with equal probability

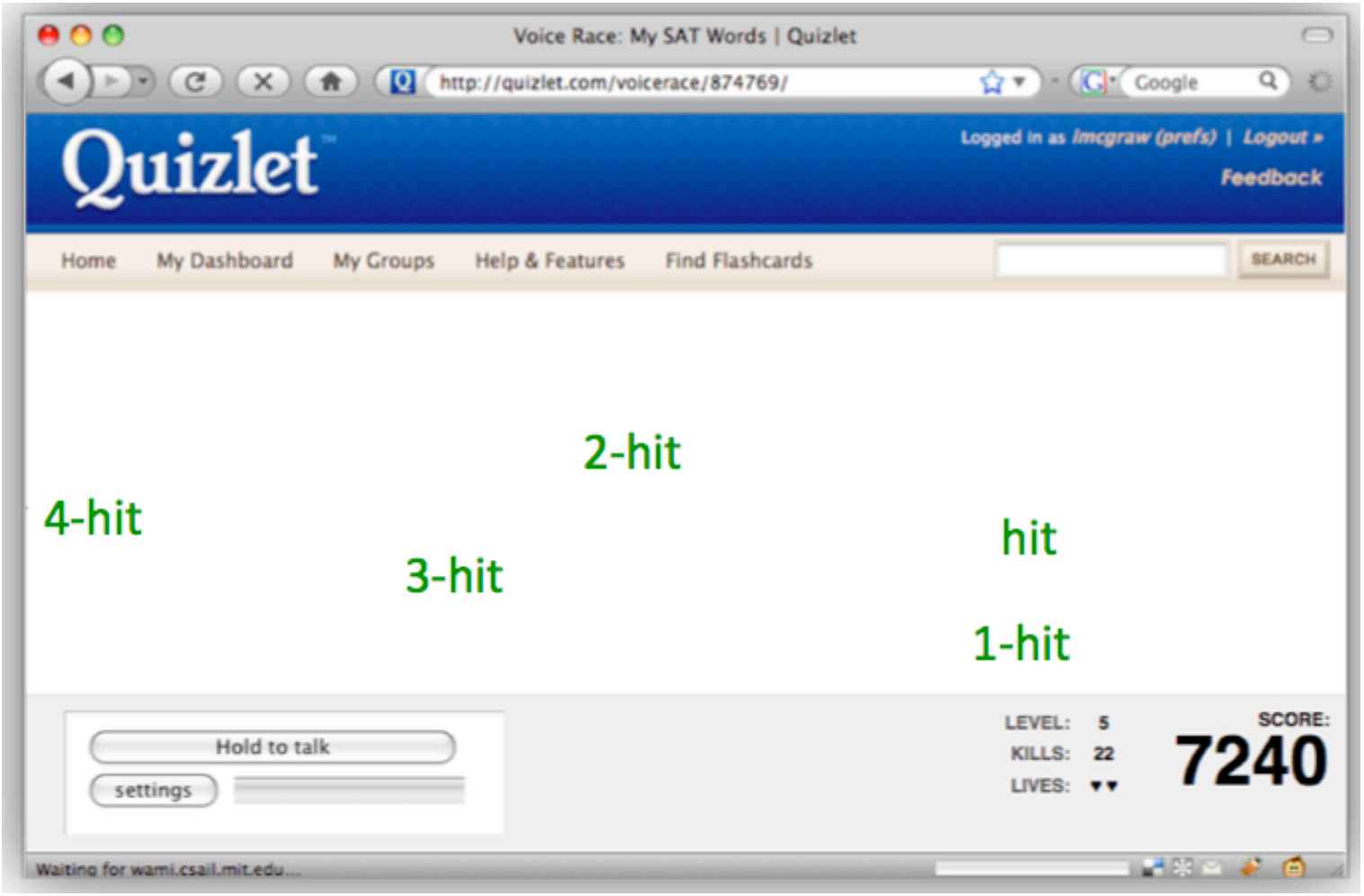
Collection results

Games Played	4,184
Utterances	55,152
Total Hours of Audio	18.7
Mean Words per Utterance	1.54
Total Distinct Phrases	26,542
Mean Category Size	53.6

Initial evaluation

- Labeled ~10000 using AMT
  - Cost $275
  - Time <24 hours
- Error rate 46.8%
  - Noisy classrooms
  - Bad microphones

Filtering by context



User tends to say the 1-hit first

Self-transcribed utterances

- The hypothesis matches the system context in a way that is unlikely to occur by chance
- Only use the utterance for semi-supervised training when the 1-hit word is the same as the ASR transcription

Recognition accuracy of utterances broken down by hit-type				
	4-hit	3-hit	2-hit	1-hit
% Accuracy	41.3	69.4	81.7	**98.5**
% of hit-data	1.8	3.4	9.0	**69.4**

Semi-supervise training

- Baseline: 46.8%
- Confidence-based filtering: 43.9%
- 1-hit based filtering: 41.2%

- Create the application in a clever way to facilitate data collection

Today topics

- VAD
- Noise reduction
  - Automatic Gain control
  - spectral subtraction
  - adaptive filtering
  - microphone array
  - Multi-condition training
- Language ID/Speaker ID/Emotion ID
  - i-vector
- Adaptation
  - VTLN
  - fMLLR
  - DNN adaptation
- Semi-supervised training and crowd sourcing
- Keyword search

Search audio content

Search audio is still mostly done by tags, word description, or titles

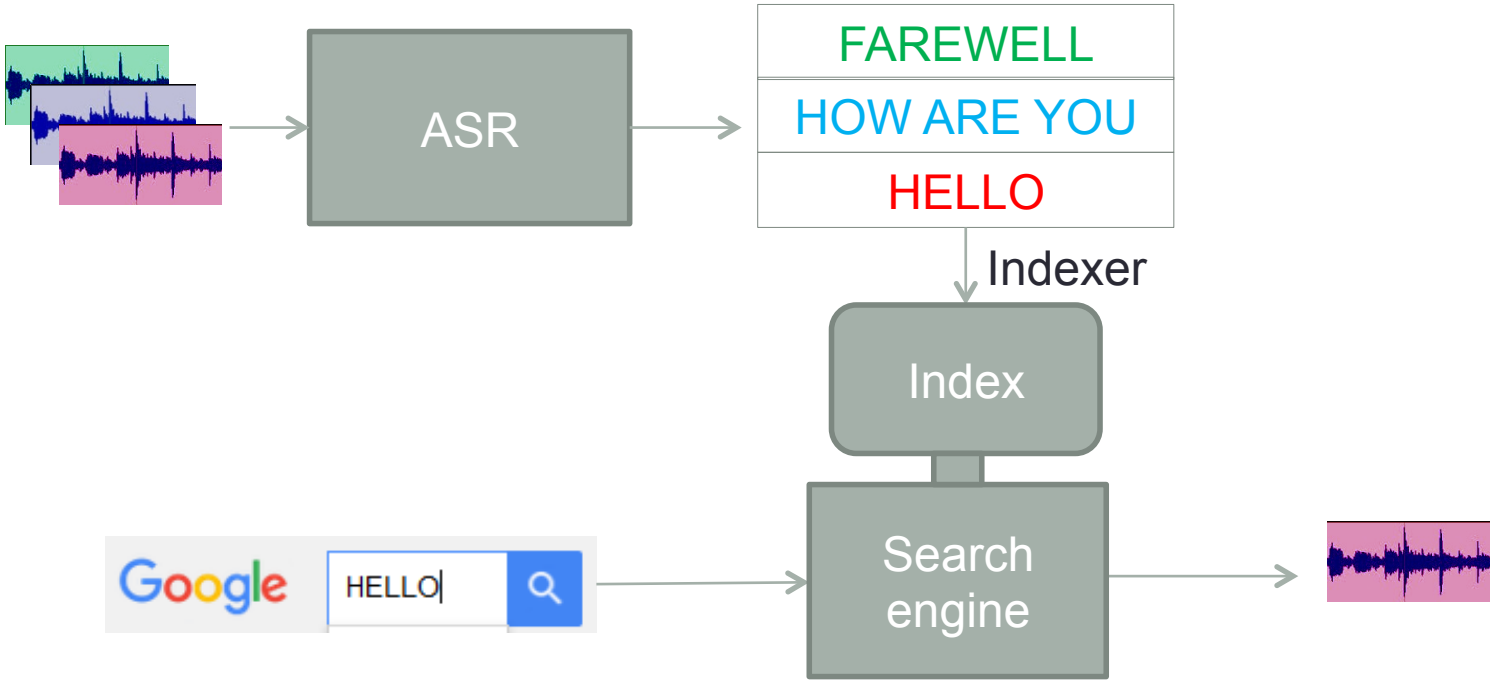
Usage for keyword search

- Searching for content of interest
  - Defense and security
  - Tagging and organizing of content

https://www.youtube.com/watch?v=IwvhdfvaJNg

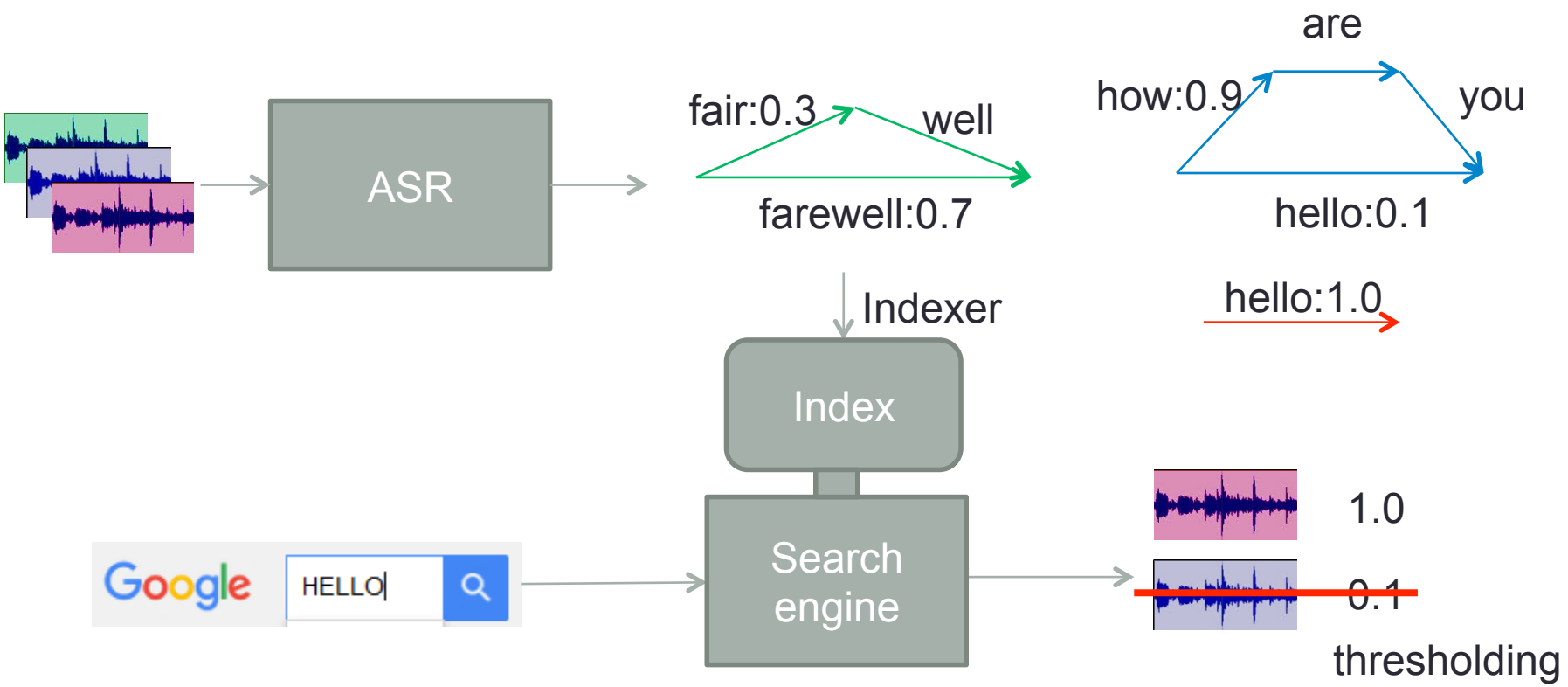
Keyword Spotting (KWS)

- Identify portions of a audio collection that contain the specified keywords
  - A keyword can be a single word or phrase



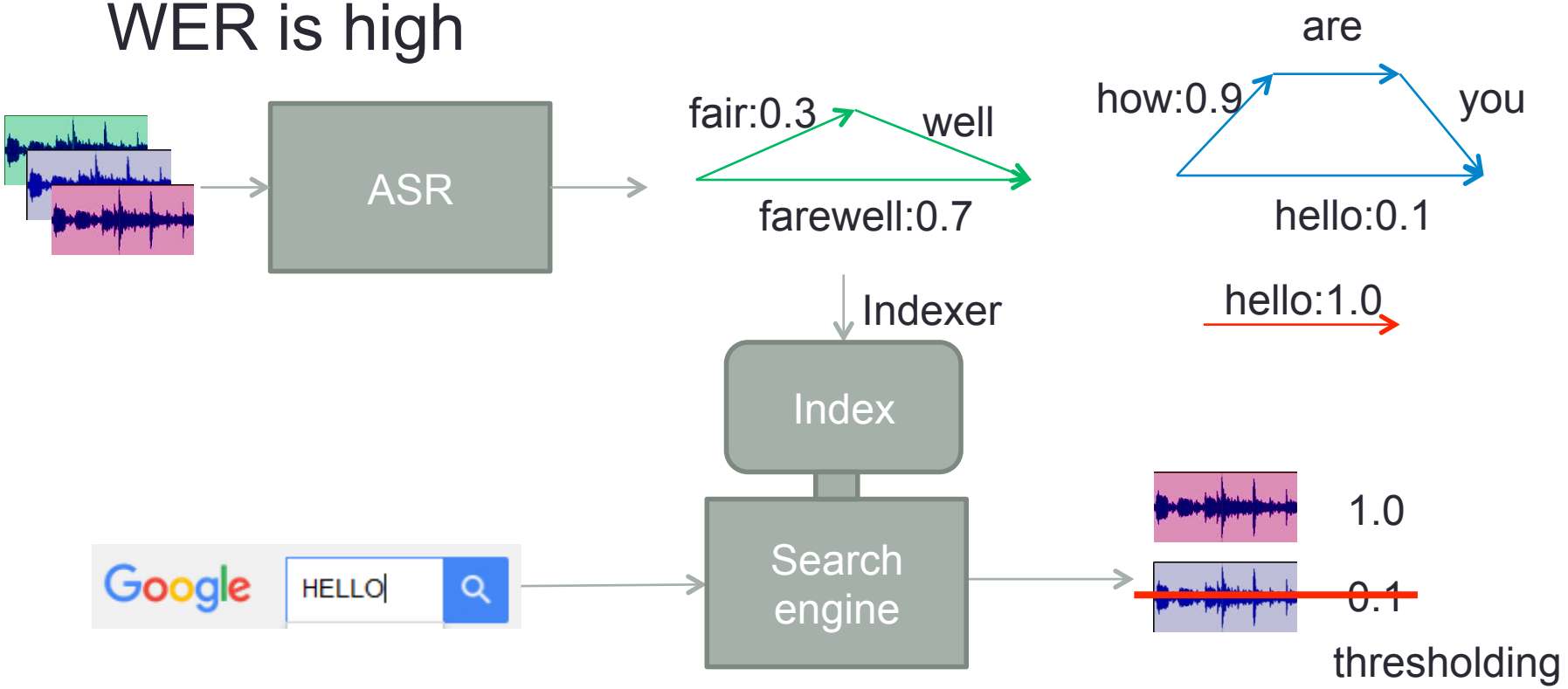
Keyword Spotting (KWS)

- Search can be done on lattices (FSTs)



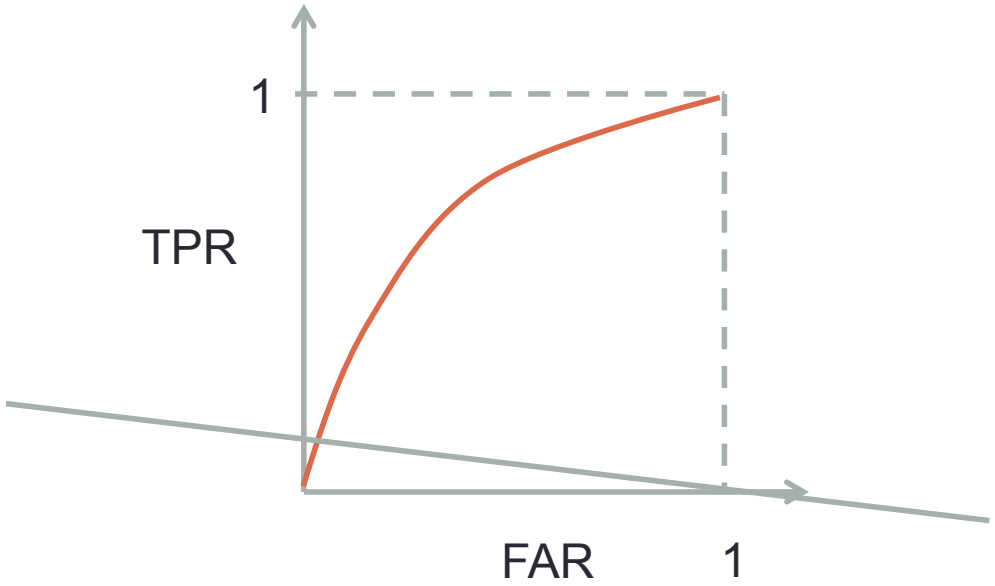
Keyword Spotting (KWS)

- Better WER *usually* means better KWS
- Keyword spotting can be robust even when the WER is high



Measuring KWS performance

- Treat as a detection problem
- Define a miss detection vs false alarm trade-off
- One standard trade-off is called Average Term Weighted Value (ATWV) with false alarm cost of 1000 times



ATWV metric

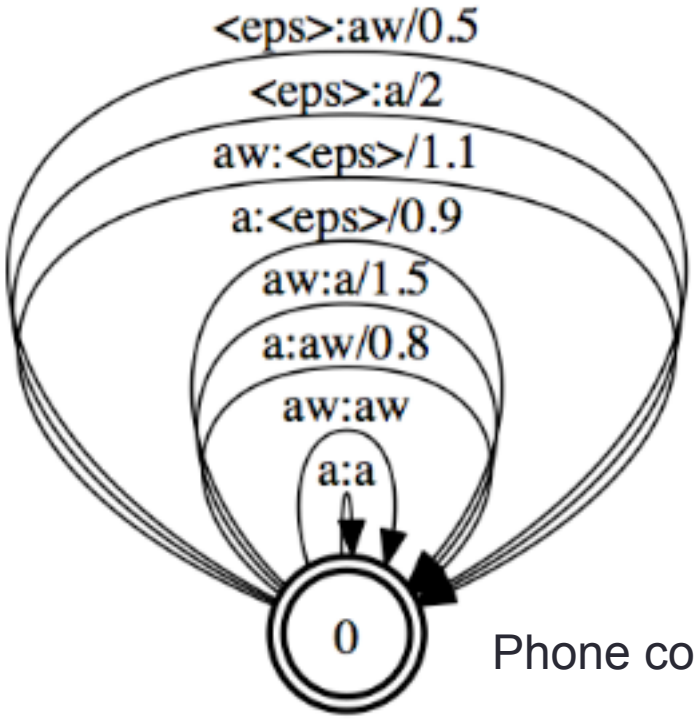
- Maximum value of 1 for perfect results
  - Can go negative if too many false alarms
- A tolerable performance is around 0.3
  - Can hit this number even with WER > 70%
  - Human will not accept these transcriptions
  - A use case for hard ASR tasks
    - Noisy
    - Large vocab
    - Little data

Two types of KWS

- Known keywords
  - Keyword is known ahead of time
  - Can add to vocabulary and boost the LM score
  - Need to run decoding after the keyword is known
  - Higher performance
- Unknown keywords
  - Keyword is given as the user wants the result
  - Problem arises when encountering OOV

Handling OOV

- Search by word proxy
  - Search the most similar sounding word in the vocabulary
    - Use FSTs to find such word



Phone confusion FST

ATWV for →	IV Kwds	OOV Kwds	All Kwds
Without Proxies	0.351	0.000	0.216
With Proxies	0.351	0.110	0.258

G. Chen, Using Proxies for OOV Keywords in the Keyword Search Task, 2013

Query expansion

- Sometimes the user does not know the exact phrase to search
- Recording about the subject but does not contain the exact keyword
- Search for related terms
  - User “Airplane”, system search “Airplane” and “Aircraft”
  - Very similar approach to text search with some modification to handle errors in ASR
  - Search by topic
  - TF-IDF (term frequency-inverse document frequency) measure terms that are very specialize in certain topics

Query expansion results

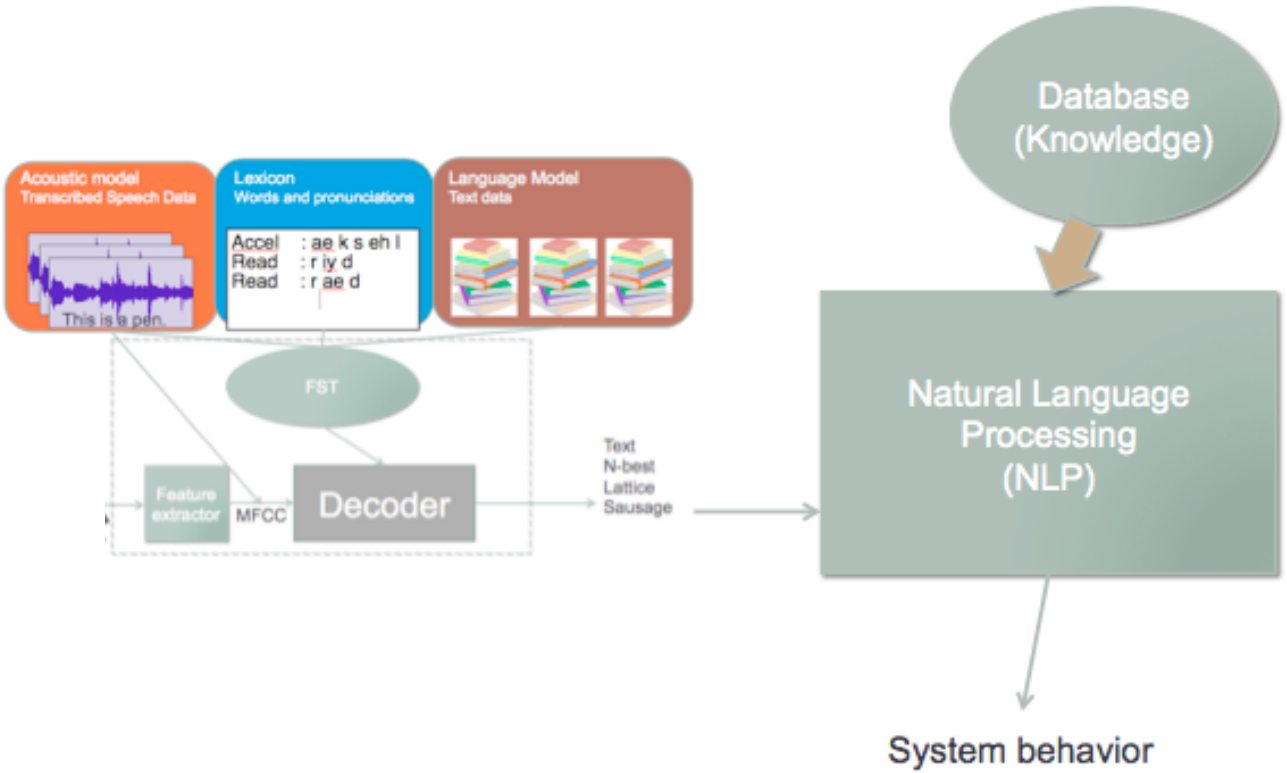
- MAP – Mean Average precision
- Search on broadcast news
- Word – expand based on similar meaning words
- Topic – expand based on similar topics that the word appear

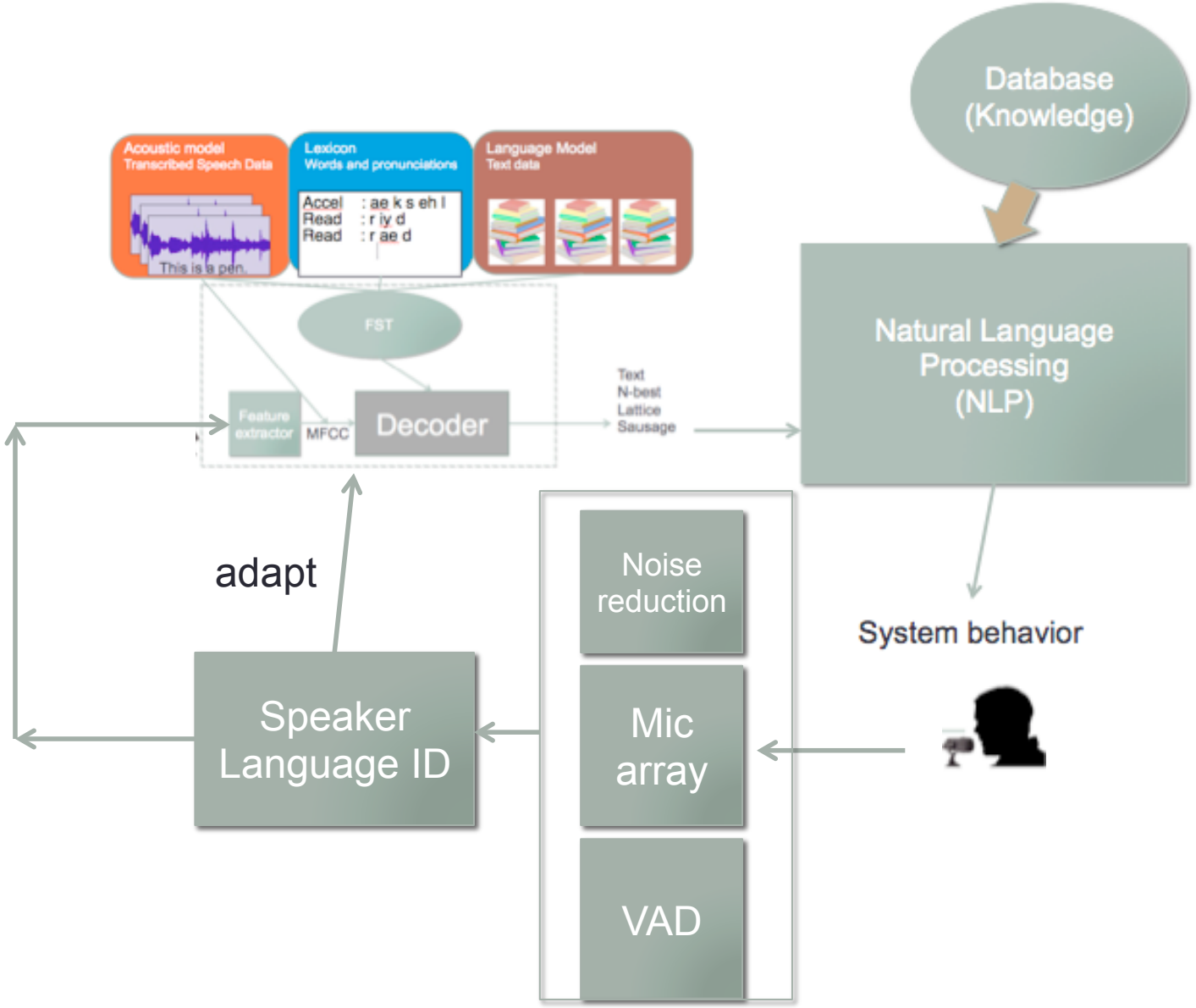
MAP	MLE	QE_{word}	QE_{topic}	$QE_{word+topic}$
Manual	53.99	57.38	17.37	61.93
1-Best	37.07	40.17	13.49	40.53
Lattice	38.09	40.48	13.21	40.79
Enhanced Lattice	40.05	41.73	13.73	41.90

T.W. Tu, Semantic query expansion and context-based discriminative term modeling for spoken document retrieval. 2012

Today topics

- VAD
  - The start of every speech application
- Noise reduction
  - Automatic Gain control
  - spectral subtraction
  - adaptive filtering
  - microphone array
  - Multi-condition training
- Language ID/Speaker ID/Emotion ID
  - Extracting a low dimensional representation
- Adaptation
  - Increase performance by matching speaker/environment
- Semi-supervised training and crowd sourcing
  - Help grows your data
- Keyword search
  - Search related terms, in meaning and pronunciation


Acoustic model
Transcribed Speech Data
This is a pen.
Lexicon
Words and pronunciations
Accel : ae k s eh l
Read : r iy d
Read : r ae d
Language Model
Text data
FST
Feature extractor
MFCC
Decoder
Text
N-best
Lattice
Sausage
Database (Knowledge)
Natural Language Processing (NLP)
System behavior


Database
(Knowledge)
Acoustic model
Transcribed Speech Data
This is a pen.
Lexicon
Words and pronunciations
Accel : ae k s eh l
Read : r iy d
Read : r ae d
Language Model
Text data
FST
Feature extractor
MFCC
Decoder
Text
N-best
Lattice
Sausage
Natural Language Processing (NLP)
adapt
Noise reduction
Mic array
VAD
Speaker Language ID
System behavior

Final words

- There are still more sound processing tasks
  - Music
  - Non-speech events (Acoustic Scene Analysis)
  - Audio-visual
  - Mind reading?

- Privacy is becoming a concern
  - Should continuous listening device be used as evidence?
  - Will privacy exist in the future?
  - The double-edge sword of technology

- ASR is continuously changing, and many companies are building ASR teams
- Smart home is the next battlefront

https://www.telegraph.co.uk/technology/2017/04/19/mark-zuckerberg-confirms-facebook-working-mind-reading-technology/

Project presentation

- 15 minutes
- Describe your app, your allowable commands
- Describe how you build the LM
- Any other tweaks you tried
  - What worked
  - What did not work
  - We wanted to know your design process
- Final performance